MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0117
De.2. **Measure Title:** Beta Blockade at Discharge
Co.1.1. **Measure Steward:** The Society of Thoracic Surgeons
De.3. **Brief Description of Measure:** Percent of patients aged 18 years and older undergoing isolated CABG who were discharged on beta blockers
1b.1. **Developer Rationale:** The use of postoperative b-blockers is now known to protect patients both at one year and long term (greater than 5 years) from death following cardiac surgery. This effect is associated with a 46 % risk reduction in death at one –year and 35% risk reduction in mortality during long-term follow-up (see Chan below). The summary of peer reviewed literature cited below supports that the utilization of beta-blocker at discharge as conferring a strong risk reduction in mortality.

- Crystal E, Connolly SJ, Sleik K, et al. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. Circulation. 2002;106(1):75-80.
- Kim MH, Deeb GM, Morady F, et al. Effect of postoperative atrial fibrillation on length of stay after cardiac surgery (The Postoperative Atrial Fibrillation in Cardiac Surgery study [PACS(2)]). Am J Cardiol. 2001;87(7):881-885.
- Maisel WH, Rawn JD, Stevenson WG. Atrial fibrillation after cardiac surgery. Ann Intern Med. 2001;135(12):1061-1073.
- Villareal RP, Hariharan R, Liu BC, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. J Am Coll Cardiol. 2004;43(5):742-748.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg. 2004;77:1137-1139.
- Charlson ME, Isom OW. Care after coronary-artery bypass surgey. N Engl J Med. 2003;348:1456-63.
- Chen J, Radford MJ, Wang Y, Marciniak TA, Krumholz HM. Are beta-blockers effective in elderly patients who undergo coronary revascularization after acute myocardial infarction? Arch Intern Med. 2000;160:947-52.
- Chan AYM, McAlister FA, Norris, CM, et al. Effect of B-Blocker use on outcomes after discharge in patients who underwent cardiac surgery. J Thorac Cardiovasc Surg. 2010;140:182-7.
- Zhang H, Yuan X, Zhang H, et al. Efficacy of long-term Beta-blocker therapy for secondary prevention of long-term outcomes after coronary artery bypass grafting surgery. Circulation 2015; 131:2194-201.
- Philip F, Blackstone E, Kapadia SR. Impact of statin and beta blocker therapy on mortality after coronary artery bypass grafting surgery. Cardiovasc Diagn Ther 2015 ; 5:8-16

S.4. **Numerator Statement:** Number of patients undergoing isolated CABG who were discharged on beta blockers
S.7. **Denominator Statement:** Patients undergoing isolated CABG
S.10. **Denominator Exclusions:** Cases are removed from the denominator if there was an in-hospital mortality or if discharge beta blocker was contraindicated.

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: May 09, 2007 **Most Recent Endorsement Date:** Jan 31, 2012

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

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

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure? X Yes**
- **Quality, Quantity and Consistency of evidence provided? X Yes**
- **Evidence graded? X Yes**

Evidence Summary and **Summary of prior review reported in 2012**

- The 2011 ACCF/AHA Guideline for Coronary Artery Bypass Graft Surgery includes the recommendation that beta blockers should be prescribed to all CABG patients without contraindications at the time of hospital discharge (Class 1C).
- During previous review, the Committee agreed the evidence, which included multiple systematic reviews, was appropriate and consistent for use of b-blockers post isolated CABG.
- The developer states that the use of postoperative b-blockers is now known to protect patients, both at one year and long term, from death following cardiac surgery.
- The summary of peer reviewed literature provided by the developer at last maintenance review and the current review supports that the utilization of beta-blocker at discharge confers a strong risk reduction in mortality.

Changes to evidence from last review

X The developer provided additional updated evidence for this measure:

- Zhang H, Yuan X, Zhang H, et al. Efficacy of long-term Beta-blocker therapy for secondary prevention of long-term outcomes after coronary artery bypass grafting surgery. Circulation 2015; 131:2194-201.
- Philip F, Blackstone E, Kapadia SR. Impact of statin and beta blocker therapy on mortality after coronary artery bypass grafting surgery. Cardiovasc Diagn Ther 2015 ; 5:8-16
- Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.

Exception to evidence N/A

The evidence includes the guideline cited above and is directionally the same compared to that for the previous NQF review.

Questions for the Committee:

- *Does the Committee agree that the evidence continues to support use of the IMA graft in patients undergoing CABG as specified in the measure? Does the Committee believe there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm: Process measure/systematice review and grading evidence (Box 3) →

Information on QQC presented (Box 4) → SR conclusion (Box 5b) → Moderate

Preliminary rating for evidence: ☐ **High** **X Moderate** ☐ **Low** ☐ **Insufficient**

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

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

- At last review, the Committee identified the measure as important and, with a performance compliance mean of 95.1 percent and median of 96.9 percent, representative of a gap for which continued performance improvement was desirable. Most recent data is drawn from two consecutive time periods (10/2013 – 9/2014 and 10/2014 – 9/2015). Mean performance during those periods was 98% in each period (1,058 participants and 139,921 operations in 2013-14 period and 1,036 participants and 139,564 operations in 2014-15). All eligible operations are included except those for which beta blocker use was contraindicated. Performance at the 10th decile ranged from 73% (2013-14) and 50% (2014-15) . Geographic distribution of participants is presented. Sample size, data and other factors provided provide sufficient detail to appreciate the gap.

Disparities

- The developer provided trends across 4 one year time periods (10/2011 – 9/2012 … 10/2014 – 9/2015) for subgroups by age, gender, race, ethnicity and insurance status. The performance ranges are presented below. In each group, performance improved over the first 3 time periods and dropped slightly in the most recent period.

Gender – Male, 98.05% - 98.58%; Female, 97.72% - 98.37%

Age Groups - <75, 98.11% - 98.62%; >=75, 97.29 – 98.10%

Race – White, 97.96% - 98.54%; Black, 98.44% - 98.64%; Other, 97.44% - 98.34%

Insurance - Age>=65, lowest performance 97.04% (Medicare+Medicaid) in earliest time period to 98.50% (Medicare w/o Medicaid/Commercial as highest in most recent period and for those in the Age<65, the low in earliest time period to high in most recent time period was 97.98% (Medicare/Medicaid) and 98.87% (Commercial/HMO). The data suggests relatively uniform high use of discharge beta blockers across all groups.

Questions for the Committee:

- *How does the Committee view the performance gap in the context of both short and long term benefit of beta blockers in the population represented in the gap and the improvement since the measure was last endorsed?*
- *How should the disparities information be factored into consideration of sociodemographic factors going forward?*
- *Assuming the Committee continues to view this measure as highly credible, reliable and valid, should the measure be considered for Inactive Endorsement with Reserve Status on the basis of performance gap?*

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

Committee pre-evaluation comments

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

1a.

- Process measure
- Strong evidence linkage and face validity
- Close to topped out, with little improvement in performance in most recent period
- Part of the CABG composite measure

- 0117 is a process measure originally endorsed in 2007, endorsed in 2012. 3 studies have been published since last endorsed to support the use of beta blockers post CABG surgery to decrease mortality.
- This is a maintenance measure. The developer provided guidelines from 2011, which were the same guidelines used in the 2012 NQF review. They also cited more recent articles from 2015 that show the effectiveness of beta blockers in CABG, particularly on long-term survival, which is a new finding. Previously, studies on reduction in mortality focused on survival after one year.

1b.

- In 2012 the performance compliance mean was 95.1%, median 96.9%. 10/2013-9/2014 and 10/2014-9/2015 mean performance was 98%, though performance in the 10th decile ranged from 73% (2013-14)to 50% (2014-15).
- Population subgroups were analyzed and suggested relatively high use of discharge beta blockers across all groups.
- The performance gap is narrowing. It was 95.1% in 2012 and 98% in 2015. This measure is close to being topped out, but there is still opportunity at 10th decile where performance ranged from 73% (2013-14) and 50% (2014-15)

1c.

- Is this measure included in the 0696 STS CABG Composite Score?

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The measure assesses the percent of patients (18 or older) who undergo isolated CABG and are discharged on beta blockers. Denominator exclusions are in-hospital mortality or beta blocker contraindication.
- The data source for the measure is the STS Adult Cardiac Surgery Database. Data are collected using the STS database collection form (version 2.81) that includes detailed items regarding a wide range of factors including procedure and discharge medications, including whether beta blocker was prescribed, not prescribed, contraindicated.
- The measure is not risk adjusted or risk stratified.
- The measure is specified for analysis at the group/practice and facility levels and intended for use in the hospital/acute care setting.
- The developer notes there have been no changes to the measure specifications since it was last endorsed.

Questions for the Committee :

- *Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- The prior review included a sample of 605 STS Adult Cardiac Surgery Database Participants who had at least 100 eligible cases for the measure and reported data to STS for all 12 months (January 1, 2009 – December 31 – 2009). Total number of potentially eligible patients was 140,573. Of these, 4,378 (0.03%) were excluded based on contraindication to the medication. Mean performance on the measure was 95.1%, with a range from 69.5%

to 100%. The issues discussed by the Committee at last review were related to patients removed from the denominator based on contraindication and clarity of the time window. Both were resolved during discussion. (*Sources: data submitted by developer at last endorsement cycle and Surgery Endorsement Maintenance report dated June 2012)*

SUMMARY OF TESTING

Reliability testing level **X Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure **X Yes** ☐ **No**

Method and Result of reliability testing

- Calculation of the measure used data from October 2014 and September 2015 from 1,036 registry participants (139,564 eligible patients) As previously, exclusions are in-hospital mortality or cases for which discharge beta blocker was contraindicated.
- Reliability was tested using a signal-to-noise approach with a hierarchical model.
- Sample size needed per participant to attain reliability of 0.50 and 0.70 was calculated. During the period October 2014 – September 2015, 95% of participants met minimum required sample size for 0.50 reliability and 76% met required sample size for 0.70 reliability.
- Overall rate of missing data is reported as 0.3%. Missing data are imputed to "no" (discharge beta blocker); participants with greater than 5% (10 of 1,048 potential participants) missing data are excluded from the measure calculation. The developer reported that 99% of participants had 4% or lower missing data.

Questions for the Committee:

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

Guidance from the Reliability Algorithm : Precise specifications (Box 1) → Empiric reliability testing (Box 2) →Testing at measure score level (Box 4) → Method described and appropriate (Box 5) → Level of confidence (Box 6)

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. **X Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

o *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- Prior maintenance review validity testing involved 40 randomly selected sites (of the 628 eligible participants that had at least 100 eligible cases for the measure and reported data for each month during January 1, 2009 – December 31, 2009), selected for audit to evaluate accuracy, consistency, and comprehensiveness of data collection and integrity. Agreement rates were calculated for 73 individual elements. Also aggregate agreement rates for each element, variable category, and overall for all categories were calculated for all sites. Testing results for discharge beta blockers – 94.5% agreement rate.

SUMMARY OF TESTING for current submission

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** X **Both**

Method of validity testing of the measure score:

☐ **Face validity only**

X **Empirical validity testing of the measure score**

Validity testing method:

- The data set of isolated CABG operations from October 2014 to September 2015 was used with few exceptions: for validity testing and the comparison of participants over time; for analysis of population disparities and for analysis on impact of exclusions.

- Data element testing was done via the STS database audit process conducted by a third party, by which participant sites are randomly selected, on an annual basis, to undergo audit of the accuracy, consistency, and comprehensiveness of data collection. In 2015, 10% of STS Adult Cardiac Surgery Database participants (107) were audited. Auditing involved reabstraction of data for 20 cases from each audited participant and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each variable, each variable category and overall.

Results:

- Face validity testing overall aggregate agreement was 96.17%.

Performance measure score testing was done using face validity and predictive validity - assessing stability of performance over time.

- Face validity was based on the fact that the measure was developed with a panel of surgeon experts and statisticians. The developer states have had near-universal acceptance of this measure by all stakeholders.
- Predictive validity was assessed using a concept of "outliers". The developers posit that stability of measure scores over time may indicate that the measure is capturing an accurate indication of provider performance. *There is some disagreement about whether stability in performance demonstrates predictive validity; some would argue that changes in performance over time are to be expected—and are, in fact, desirable—as the result of quality improvement interventions. NQF guidance suggests that predictive validity should compare measure results to another measure of the same construct or to a different outcome measure.*
- Participants were placed into three groups – low performance (95% exact binomial confidence interval of event rate entirely below population average), high performance (95% confidence interval entirely above 1) and mid performance (remaining participants). Predictive validity analysis was restricted to 1,012 participants that received the measure in both time periods (10/2013 – 9/2014 and 10/2014 – 9/2015) To assess impact of the beta blocker contraindication, the distribution of the measure with and without the exclusion was computed.
- Aggregated proportion of patients receiving beta blocker at discharge in the later period (10/2014 – 9/2015) was also calculated and is reported as 94.3% (low performance), 98.8 (mid performance), and 99.6% (high performance) with the conclusion that the measure reflects the proportion of patients discharged on beta blockers and that the past measure can be used to predict future performance.
- Validity testing results showed that STS registry participants in low, middle and high groupings for use of beta blocker at discharge in one time period (10/2013 – 9/2014) had correspondingly low, middle, and high beta blocker at discharge rates in the following period (10/2014 – 9/2015).
- Of high performers in the earlier period, 76.1% were high performers in the second period. Only 5.2% of mid performers in the early period became high performers in the second period. The developer then concludes that a consumer may expect that a high or low performer will likely be the same or become average in near future and a mid-performer is likely to remain average in performance on the measure.

Questions for the Committee:

- *Do the results indicate there is room for improvement?*
- *Do the results allow conclusions about quality that can be translated to meaningful interpretation of high, mid and low performance by users?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The exclusions to the measure are in-hospital death and contraindication to use of beta blockers. The developer has calculated the distribution of the participant-specific observed proportion of patients receiving beta blockers with and without the exclusion for the 10/2014 – 9/2015 time period and shown the percent change across each of the three performance groups to demonstrate its effect on measure results.
- With respect to the beta blocker contraindication exclusion criteria, proportion of patients receiving beta blockers changed as follows: low performance participants with exclusion, 9.1%; without, 9.9%; mid performance 82.9% with exclusion; without, 80.1%; high performance 8.0%; without, 9.9%. The developer reports that STS database participants performing better and worse than the STS average has remained similar over the two time periods, 10/2013 – 9/2014 and 10/2014 – 9/2015, with more than 80% having performance indistinguishable from the STS average.

Questions for the Committee:

- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** **X None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- As noted in earlier sections, the developer has presented information about low, mid, and high performance participants.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed. The measure uses a single data source and has one set of specifications.

2b7. Missing Data

- As noted earlier, the overall rate of missing data is reported as 0.3%. Missing data are imputed to "no" (discharge beta blocker); participants with greater than 5% (10 of 1,048 potential participants) missing data are excluded from the measure calculation. The developer reported that 99% of participants had 4% or lower missing data.

Guidance from the Validity Algorithm: Specifications consistent with evidence (Box 1) → Threats to validity (Box 2) → Empirical validity testing (Box 3) → Testing with performance measure scores (Box 6) → Method described and appropriate (Box 7) → Confidence that scores are indicator of quality (Box 8) Of note, patient level data elements also tested.

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

Committee pre-evaluation comments

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

2a.

- Beta blocker prescribed at discharge should be clearly documented for abstraction. The data elements are not risk adjusted and denominator exclusions include in-hospital mortality and contraindication to beta blocker.

- The measure can be consistently implemented.
- The measure has been assessed in the past as having appropriate specificity. There are two points that I would like clarification on, however, 1) is a ""contraindication"" simply the presence of a statement, at the physician's discretion, that says a beta blocker is contraindicated or does it require that a defined set of accepted contraindicatory criteria are met? 2) beta blocker dose is not specified. Is this also left to provider discretion? Are the benefits of giving beta blockers the same at all dosages or is there a narrow range of widely accepted dosing for beta blockers?

2a2.

- Reliability testing was conducted at the measure score using both the data element and score levels using signal-to-noise approach with a hierarchical model.
- 10/2014-9/2015: 95% of participants met the minimum required sample size for 0.50 reliability and 76% met required sample size for 0.70 reliability. 0.3% missing data, ""no"", was imputed as 0. Greater than 5% missing data were excluded. 99% of participants had 4% or lower missing data."
- Data provided by the developer shows the measure to be highly reliable. During the period October 2014 – September 2015, 95% of participants met minimum required sample size for 0.50 reliability and 76% met required sample size for 0.70 reliability.

2b1.

- Specifications were consistent with the evidence provided.
- The process for evaluating validity appears to be thorough and free of confounding factors.

2b2.

- Validity testing was adequate in scope. Empirical validity testing was conduced at both the measure score and 73 individual elements tested against a gold standard.
- Face validity testing overall aggregate agreement was 96.17%.
- During the last discussion, the use of predictive validity testing was called into question. Other methods of validity testing were also employed however and the measure does appear to have adequate validity.

2b3.

- Missing data does not constitute and threat to measure validity.
- Missing data is reported at 0.3%. Missing data, ""no"", was imputed as 0. Greater than 5% missing data were excluded. 99% of participants had 4% or lower missing data.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.

- Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants. Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.

- There are no additional costs for data collection specific to the measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

Questions for the Committee:

- *Is the effort and cost associated with abstracting the required data elements appropriate to the value of the*

measure?

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

Committee pre-evaluation comments

Criteria 3: Feasibility

- Required data elements are generated by healthcare personnel during documentation of care.
- If manual chart extraction is required, additional facility cost is incurred.
- This is a simple and highly feasible metric to implement, document, and track.

Criterion 4: Usability and Use

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

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

Current uses of the measure

Publicly reported? X **Yes** ☐ **No**

Current use in an accountability program? X **Yes** ☐ **No**

OR

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

Accountability program details

- One of 11 component measures of the STS CABG Composite Score reported by STS Public Reporting Online and Consumer Reports Health. Some 49.8% of the STS Adult Cardiac Surgery Database's 1,100 participants are voluntarily enrolled in the public reporting program. STS analyses indicate that the STS database includes more than 90% of cardiothoracic programs in the US.
- Used for QI with benchmarking (involves external benchmarking with multiple organizations).
- QI internal to the organization

Improvement results

- Aggregate proportion of eligible patients receiving beta blockade at discharge has ranged from a low of 97.96% for the period 10/2011 – 9/2012 to a high of 98.68% in 10/2013 – 9-2014. A slight drop to 98.53% occurred in 10/2014 – 9/2015.

Unexpected findings (positive or negative) during implementation: None identified

Potential harms:

- The developer reports it is not aware of any negative unintended consequences, noting that all public reporting has potential for such things as gaming and risk aversion. The developer attempts to control gaming though its audit process and risk aversion though a methodology that adjusts expected risk for providers who care for sicker patients.

Feedback :

- At the time of endorsement reported in the June 2012 Surgery Endorsement Maintenance report, public comment noted that the measure was considered to be topped out with a mean value at 95.5%. At the time, the Committee stated that the distribution of measure values indicate opportunity for improvement. With respect to combining the measures, the Committee agreed that the difference in denominators and use in the STS CABG Composite Score are acceptable reasons to continue the measure as a standalone.

***Questions for the Committee**:*

- o *Does the measure continue to be useful for improvement given the sustained high level of performance with little change since last review?*
- o *How does the use of the measure as part of an 11 component CABG composite score affect Committee view of its ongoing value?*

Preliminary rating for usability and use: ☐ **High** ☐ **Moderate** **X Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Publically reported as one of 11 component measures of the STS CABG Composite Score reported by STS Public Reporting Online and Consumer Health Reports for more than 90% of US cardiothoracic programs.
- Post operative beta blockade may not be the best management choice for all patients. If the hospital stay is short, there may not be adequate time to assess pharmacologic management to assure safe medication management post discharge.
- Measures are reported through the STS database which has a large-scale program aimed at improving the quality of cardiothoracic surgery nationwide.

Criterion 5: Related and Competing Measures

Harmonization and Related or competing measures

Measures 0117 and 0127 are STS measures of beta blocker use that are harmonized. CMS Measure #0284, Surgery Patients on Beta-Blocker Therapy Prior to Arrival Who Received a Beta-Blocker During the Perioperative Period is a similar, though more broad, measure, based on administrative claims data for use at the facility or regional levels. The CMS measure has a greater number of exclusions and extensive data collection algorithm. Uses of the measures differ. Other related measures include the STS measures that comprise the STS CABG Composite Score

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0117

Measure Title: Beta Blockade at Discharge

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 0696 STS CABG Composite Score

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☐ Health outcome:

☐Patient-reported outcome (PRO): 42T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 42T

☒ Process: beta blocker at discharge

☐ Structure: 42T

☐ Other: 42T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The summary of peer reviewed literature cited below supports that the utilization of beta-blocker at discharge as conferring a strong risk reduction in mortality. In addition, CABG is a frequently performed procedure, and a large number of patients undergo CABG yearly in the US. The development of post-operative atrial fibrillation consumes excess resources.

- Crystal E, Connolly SJ, Sleik K, et al. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. *Circulation.* 2002;106(1):75-80.
- Kim MH, Deeb GM, Morady F, et al. Effect of postoperative atrial fibrillation on length of stay after cardiac surgery (The Postoperative Atrial Fibrillation in Cardiac Surgery study [PACS(2)]). *Am J Cardiol.* 2001;87(7):881-885.
- Maisel WH, Rawn JD, Stevenson WG. Atrial fibrillation after cardiac surgery. *Ann Intern Med.* 2001;135(12):1061-1073.
- Villareal RP, Hariharan R, Liu BC, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. *J Am Coll Cardiol.* 2004;43(5):742-748.

- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg.* 2004;77:1137-1139.
- Charlson ME, Isom OW. Care after coronary-artery bypass surgey. *N Engl J Med*. 2003;348:1456-63.
- Chen J, Radford MJ, Wang Y, Marciniak TA, Krumholz HM. Are beta-blockers effective in elderly patients who undergo coronary revascularization after acute myocardial infarction? *Arch Intern Med*. 2000;160:947-52.
- Chan AYM, McAlister FA, Norris, CM, et al. Effect of B-Blocker use on outcomes after discharge in patients who underwent cardiac surgery. *J Thorac Cardiovasc Surg*. 2010;140:182-7.
- Zhang H, Yuan X, Zhang H, et al. Efficacy of long-term Beta-blocker therapy for secondary prevention of long-term outcomes after coronary artery bypass grafting surgery. Circulation 2015; 131:2194-201.
- Philip F, Blackstone E, Kapadia SR. Impact of statin and beta blocker therapy on mortality after coronary artery bypass grafting surgery. Cardiovasc Diagn Ther 2015 ; 5:8-16.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.

http://circ.ahajournals.org/content/124/23/e652

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page e152

4.5. Perioperative Beta Blockers: Recommendations

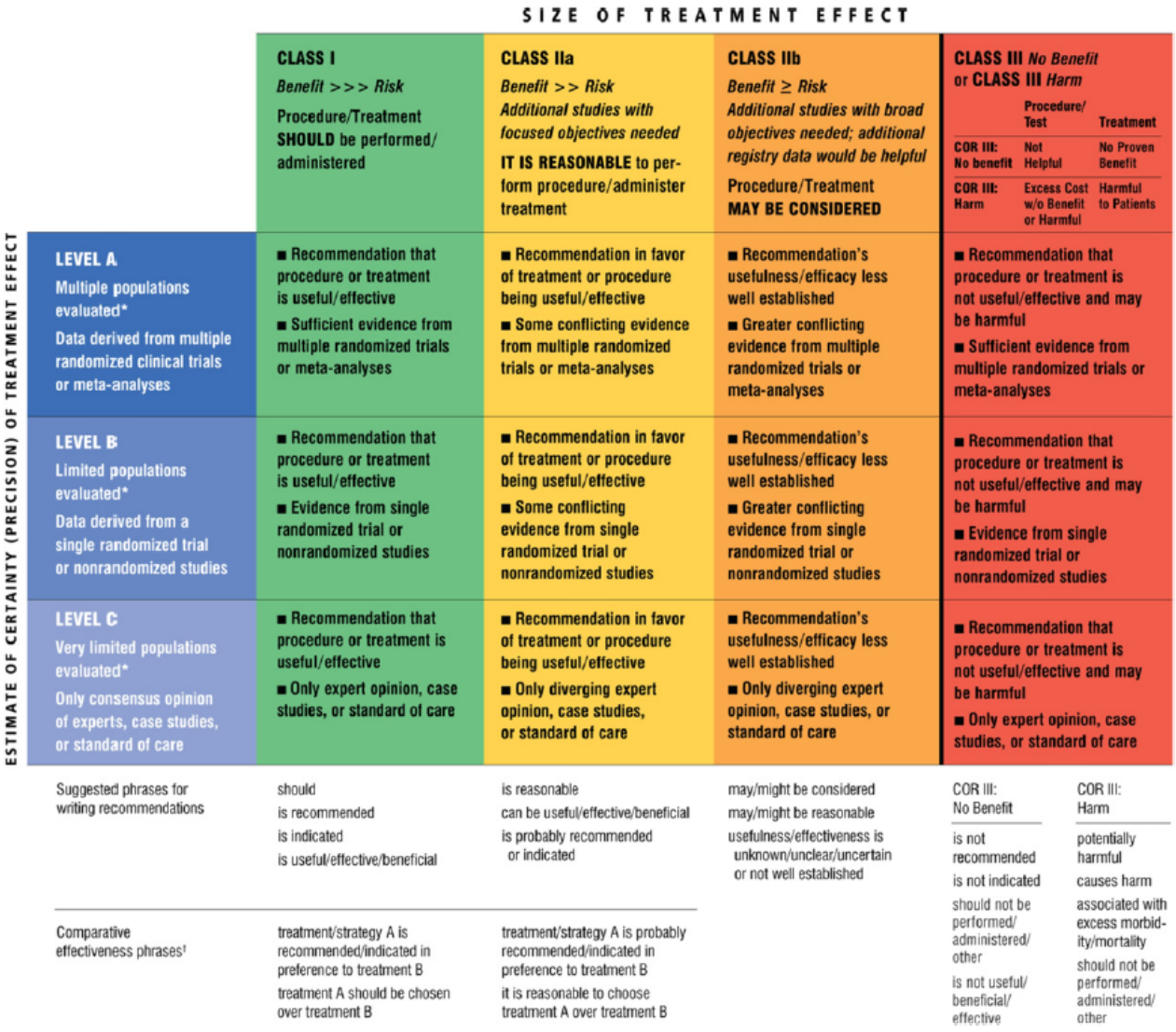
Class I Recommendation

Beta blockers should be prescribed to all CABG patients without contraindications at the time of hospital discharge. (Level of Evidence: C)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Class 1 Level C. Recommendation that procedure or treatment is useful/effective. Only expert opinion, case studies, or standard of care.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)



SIZE OF TREATMENT EFFECT

ESTIMATE OF CERTAINTY (PRECISION) OF TREATMENT EFFECT	CLASS I *Benefit >>> Risk* Procedure/Treatment **SHOULD** be performed/ administered	CLASS IIa *Benefit >> Risk* *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	CLASS IIb *Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	CLASS III *No Benefit* or CLASS III *Harm* Procedure/Test / Treatment COR III: No benefit — Not Helpful / No Proven Benefit COR III: Harm — Excess Cost w/o Benefit or Harmful / Harmful to Patients
LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical trials or meta-analyses	■ Recommendation that procedure or treatment is useful/effective ■ Sufficient evidence from multiple randomized trials or meta-analyses	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Sufficient evidence from multiple randomized trials or meta-analyses
LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is useful/effective ■ Evidence from single randomized trial or nonrandomized studies	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Evidence from single randomized trial or nonrandomized studies
LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	■ Recommendation that procedure or treatment is useful/effective ■ Only expert opinion, case studies, or standard of care	■ Recommendation in favor of treatment or procedure being useful/effective ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation's usefulness/efficacy less well established ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Only expert opinion, case studies, or standard of care
Suggested phrases for writing recommendations	should is recommended is indicated is useful/effective/beneficial	is reasonable can be useful/effective/beneficial is probably recommended or indicated	may/might be considered may/might be reasonable usefulness/effectiveness is unknown/unclear/uncertain or not well established	COR III: No Benefit: is not recommended; is not indicated; should not be performed/ administered/ other; is not useful/ beneficial/ effective COR III: Harm: potentially harmful; causes harm; associated with excess morbidity/mortality; should not be performed/ administered/ other
Comparative effectiveness phrases†	treatment/strategy A is recommended/indicated in preference to treatment B treatment A should be chosen over treatment B	treatment/strategy A is probably recommended/indicated in preference to treatment B it is reasonable to choose treatment A over treatment B		

A recommendation with Level of Evidence B or C does not imply that the recommendation is weak. Many important clinical questions addressed in the guidelines do not lend themselves to clinical trials. Although randomized trials are unavailable, there may be a very clear clinical consensus that a particular test or therapy is useful or effective.

*Data available from clinical trials or registries about the usefulness/efficacy in different subpopulations, such as sex, age, history of diabetes, history of prior myocardial infarction, history of heart failure, and prior aspirin use. †For comparative effectiveness recommendations (Class I and IIa; Level of Evidence A and B only), studies that support the use of comparator verbs should involve direct comparisons of the treatments or strategies being evaluated.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

ACCF/AHA Task Force on Practice Guidelines. Methodologies and Policies from the ACCF/AHA Task Force on Practice Guideline. June 2010.

http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf and

http://circ.ahajournals.org/site/manual/index.xhtml

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Information relevant to this section is provided in detail in the previous sections and the referenced guideline. Please see Appendix.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Improved survival; some observational analyses reported that discharge beta blockers were effective in those with perioperative myocardial ischemia or elderly subjects with heart failure.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

See 1a.7

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.7

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 42T

See 1a.7

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

See 1a.7

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

See 1a.7

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

See 1a.7

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

See 1a.7

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

See 1a.7

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1a._Evidence_-_0117_Beta_Blockade_at_Discharge-636008098416004445.docx

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
The use of postoperative b-blockers is now known to protect patients both at one year and long term (greater than 5 years) from death following cardiac surgery. This effect is associated with a 46 % risk reduction in death at one –year and 35% risk reduction in mortality during long-term follow-up (see Chan below). The summary of peer reviewed literature cited below supports that the utilization of beta-blocker at discharge as conferring a strong risk reduction in mortality.

- Crystal E, Connolly SJ, Sleik K, et al. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. Circulation. 2002;106(1):75-80.
- Kim MH, Deeb GM, Morady F, et al. Effect of postoperative atrial fibrillation on length of stay after cardiac surgery (The Postoperative Atrial Fibrillation in Cardiac Surgery study [PACS(2)]). Am J Cardiol. 2001;87(7):881-885.
- Maisel WH, Rawn JD, Stevenson WG. Atrial fibrillation after cardiac surgery. Ann Intern Med. 2001;135(12):1061-1073.
- Villareal RP, Hariharan R, Liu BC, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. J Am Coll Cardiol. 2004;43(5):742-748.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg. 2004;77:1137-1139.
- Charlson ME, Isom OW. Care after coronary-artery bypass surgey. N Engl J Med. 2003;348:1456-63.
- Chen J, Radford MJ, Wang Y, Marciniak TA, Krumholz HM. Are beta-blockers effective in elderly patients who undergo coronary revascularization after acute myocardial infarction? Arch Intern Med. 2000;160:947-52.
- Chan AYM, McAlister FA, Norris, CM, et al. Effect of B-Blocker use on outcomes after discharge in patients who underwent cardiac surgery. J Thorac Cardiovasc Surg. 2010;140:182-7.
- Zhang H, Yuan X, Zhang H, et al. Efficacy of long-term Beta-blocker therapy for secondary prevention of long-term outcomes after coronary artery bypass grafting surgery. Circulation 2015; 131:2194-201.
- Philip F, Blackstone E, Kapadia SR. Impact of statin and beta blocker therapy on mortality after coronary artery bypass grafting surgery. Cardiovasc Diagn Ther 2015 ; 5:8-16

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data*

source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
See Appendix

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

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
CABG is a frequently performed procedure, and a large number of patients undergo CABG yearly in the US. The development of post-operative atrial fibrillation consumes excess resources. The summary of peer reviewed literature cited below supports that the utilization of beta-blocker at discharge as conferring a strong risk reduction in mortality.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

- Crystal E, Connolly SJ, Sleik K, et al. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. Circulation. 2002;106(1):75-80.
- Kim MH, Deeb GM, Morady F, et al. Effect of postoperative atrial fibrillation on length of stay after cardiac surgery (The Postoperative Atrial Fibrillation in Cardiac Surgery study [PACS(2)]). Am J Cardiol. 2001;87(7):881-885.
- Maisel WH, Rawn JD, Stevenson WG. Atrial fibrillation after cardiac surgery. Ann Intern Med. 2001;135(12):1061-1073.
- Villareal RP, Hariharan R, Liu BC, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. J Am Coll Cardiol. 2004;43(5):742-748.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg. 2004;77:1137-1139.
- Charlson ME, Isom OW. Care after coronary-artery bypass surgey. N Engl J Med. 2003;348:1456-63.
- Chen J, Radford MJ, Wang Y, Marciniak TA, Krumholz HM. Are beta-blockers effective in elderly patients who undergo coronary revascularization after acute myocardial infarction? Arch Intern Med. 2000;160:947-52.
- Chan AYM, McAlister FA, Norris, CM, et al. Effect of B-Blocker use on outcomes after discharge in patients who underwent cardiac surgery. J Thorac Cardiovasc Surg. 2010;140:182-7.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;
http://www.sts.org/sites/default/files/documents/STSAdultCVDataSpecificationsV2_81.pdf

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
None

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients undergoing isolated CABG who were discharged on beta blockers

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Denominator – 12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of isolated CABG procedures in which discharge beta blockers [DCBeta (STS Adult Cardiac Surgery Database Version 2.81)] is marked "yes"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*

Patients undergoing isolated CABG

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Number of isolated CABG procedures excluding cases with an in-hospital mortality or cases for which discharge beta blocker use was contraindicated. The SQL code used to create the function used to identify cardiac procedures is provided in the Appendix.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Cases are removed from the denominator if there was an in-hospital mortality or if discharge beta blocker was contraindicated.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Mortality Discharge Status (MtDCStat), Mortality Date (MtDate), and Discharge Date (DischDt) indicate an in-hospital mortality; discharge beta blocker (DCBeta) marked as "Contraindicated"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
The source fields required by the beta blockade at discharge measure had only 0.3% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with greater than 5% missing data are excluded from the calculation of the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database Version 2.81

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
2.1_Testing_-_0117_Beta_Blockade_at_Discharge.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0117
Measure Title: Beta Blockade at Discharge
Date of Submission: 6/5/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process

☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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). [13]

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data*

specified and intended for measure implementation. ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 42T	☐ other: 42T

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

STS Adult Cardiac Surgery Database (ACSD) Version 2.81

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

October 2014 – September 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 42T	☐ other: 42T

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

The calculation of the beta blockade at discharge measure of the 12 months from October 2014 to September 2015 used 139,564 operations from 1,036 STS ACSD participants.

Distribution of participant sample sizes (denominator), and observed proportion of patients receiving the measure (numerator/denominator)

Stat	N	% Beta Blockade at discharge

N	1036.0	1036.0
Mean	134.7	98.0
STD	107.2	3.9
IQR	115.0	2.4
0%	2.0	50.0
10%	37.0	94.3
20%	54.0	97.0
30%	71.0	98.1
40%	85.0	98.8
50%	103.5	99.3
60%	128.0	100.0
70%	156.0	100.0
80%	201.0	100.0
90%	268.5	100.0
100%	844.0	100.0

Distribution of participants by geographic regions

REGION	
Midwest	296
Northeast	136
South	389
West	215

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

All eligible isolated operations were included except cases with an in-hospital mortality or cases for which discharge beta blocker use was contraindicated.

	Effects	**Overall N=139564**
Age (years)	Median (IQR)	65.0 (58.0, 72.0)
	Missing	0 (0.0%)
Sex	Male	105,326 (75.5%)
	Female	34,176 (24.5%)
	Missing	62 (0.0%)
Race - Asian	No	132,261 (94.8%)
	Yes	4,294 (3.1%)
	Missing	3,009 (2.2%)
Race - Black / African American	No	126,041 (90.3%)
	Yes	10,517 (7.5%)
	Missing	3,006 (2.2%)
Race - White	No	20,822 (14.9%)
	Yes	115,801 (83.0%)
	Missing	2,941 (2.1%)
Race - American Indian / Alaskan Native	No	135,676 (97.2%)
	Yes	882 (0.6%)
	Missing	3,006 (2.2%)
Race - Other	No	131,683 (94.4%)
	Yes	4,495 (3.2%)

	Effects	Overall N=139564
	Missing	3,386 (2.4%)
Native Hawaiian / Pacific Islander	No	135,854 (97.3%)
	Yes	641 (0.5%)
	Missing	3,069 (2.2%)
Hispanic or Latino Ethnicity	No	122,462 (87.7%)
	Yes	9,839 (7.0%)
	Missing	7,263 (5.2%)
Insurance: Younger than 65	Medicare/Medicaid	17,491 (27.2%)
	Commercial/HMO	38,339 (59.6%)
	None/Self Paid	5,085 (7.9%)
	Other	3,360 (5.2%)
Insurance: 65 or Older	Medicare+Medicaid	4,763 (6.3%)
	Medicare+Commercial without Medicaid	41,526 (55.2%)
	Medicare without Medicaid/Commercial	29,000 (38.5%)
Region	NORTHEAST	22,351 (16.0%)
	SOUTH	60,956 (43.7%)
	MIDWEST	34,154 (24.5%)
	WEST	22,103 (15.8%)
Body Surface Area (m)	<1.5	1,766 (1.3%)
	>=1.5 and <1.75	16,575 (11.9%)
	>=1.75 and <2	47,745 (34.2%)
	>=2	73,429 (52.6%)
	Missing	49 (0.0%)
Diabetes	No Diabetes	72,185 (51.7%)
	Diabetes - Noninsulin	41,308 (29.6%)
	Diabetes - Insulin	24,619 (17.6%)
	Diabetes - Other	369 (0.3%)
	Diabetes - Missing Treatment	783 (0.6%)
	Missing	300 (0.2%)
Hypertension	No	15,261 (10.9%)
	Yes	124,016 (88.9%)
	Missing	287 (0.2%)
Renal Function	Creatinine <1 mg/dL	67,662 (48.5%)
	Creatinine 1-1.5 mg/dL	56,437 (40.4%)
	Creatinine 1.5-2 mg/dL	8,187 (5.9%)
	Creatinine 2-2.5 mg/dL	1,742 (1.2%)
	Creatinine >2.5 mg/dL	1,317 (0.9%)
	Dialysis	3,921 (2.8%)
	Missing	298 (0.2%)
Dyslipidemia	No	16,601 (11.9%)
	Yes	122,356 (87.7%)
	Missing	607 (0.4%)
Chronic Lung Disease (CLD)	None	100,751 (72.2%)
	Mild	14,875 (10.7%)
	Moderate	6,713 (4.8%)
	Severe	5,735 (4.1%)

	Effects	Overall N=139564
	5	6,864 (4.9%)
	Missing	4,626 (3.3%)
Peripheral Vascular Disease (PVD)	No	119,135 (85.4%)
	Yes	19,529 (14.0%)
	Missing	900 (0.6%)
Cerebrovascular Disease (CVD)	No CVD	111,622 (80.0%)
	CVD-NO CVA	27,942 (20.0%)
Endocarditis	No Endocarditis	139,331 (99.8%)
	Treated Endocarditis	61 (0.0%)
	Active Endocarditis	8 (0.0%)
	Endocarditis - Missing Type	7 (0.0%)
	Missing	157 (0.1%)
Acuity Status	Elective	52,969 (38.0%)
	Urgent	80,674 (57.8%)
	Emergent	5,745 (4.1%)
	Emergent Salvage	156 (0.1%)
	Missing	20 (0.0%)
Myocardial Infarction	No Prior MI	65,332 (46.8%)
	MI >21 days	26,411 (18.9%)
	MI 8-21 days	6,673 (4.8%)
	MI 1-7 days	34,686 (24.9%)
	MI 6-24 hrs	3,285 (2.4%)
	MI <= 6 hrs	1,679 (1.2%)
	MI - Missing Timing	351 (0.3%)
	Missing	1,147 (0.8%)
Cardiogenic Shock	No	137,887 (98.8%)
	Yes	1,627 (1.2%)
	Missing	50 (0.0%)
Preop IABP	No	129,589 (92.9%)
	Yes	9,801 (7.0%)
	Missing	174 (0.1%)
Congestive Heart Failure	No CHF	111,996 (80.2%)
	CHF NYHA-I	2,314 (1.7%)
	CHF NYHA-II	8,025 (5.8%)
	CHF NYHA-III	9,566 (6.9%)
	CHF NYHA-IV	5,513 (4.0%)
	CHF Missing NYHA	926 (0.7%)
	Missing	1,224 (0.9%)
Number of Diseased Coronary Vessels	None	130 (0.1%)
	One	5,844 (4.2%)
	Two	27,143 (19.4%)
	Three	105,488 (75.6%)
	Missing	959 (0.7%)
Left Main Disease > 50%	No	46,995 (33.7%)
	Yes	44,312 (31.8%)
	Missing	48,257 (34.6%)

	Effects	Overall N=139564
Ejection Fraction (%)	Median (IQR)	55.0 (45.0, 60.0)
	Missing	4,263 (3.1%)
Aortic Stenosis	No	132,712 (95.1%)
	Yes	4,129 (3.0%)
	Missing	2,723 (2.0%)
Mitral Stenosis	No	136,109 (97.5%)
	Yes	687 (0.5%)
	Missing	2,768 (2.0%)
Tricuspid Stenosis	No	136,229 (97.6%)
	Yes	89 (0.1%)
	Missing	3,246 (2.3%)
Pulmonic Stenosis	No	134,987 (96.7%)
	Yes	29 (0.0%)
	Missing	4,548 (3.3%)
Aortic Insufficiency	None	90,164 (64.6%)
	Trivial	13,336 (9.6%)
	Mild	10,506 (7.5%)
	Moderate	2,061 (1.5%)
	Severe	87 (0.1%)
	N/A or Not Documented	22,255 (15.9%)
	Missing	1,155 (0.8%)
Mitral Insufficiency	None	43,978 (31.5%)
	Trivial	33,973 (24.3%)
	Mild	32,654 (23.4%)
	Moderate	8,504 (6.1%)
	Severe	601 (0.4%)
	N/A or Not Documented	18,915 (13.6%)
	Missing	939 (0.7%)
Tricuspid Insufficiency	None	46,358 (33.2%)
	Trivial	39,643 (28.4%)
	Mild	25,503 (18.3%)
	Moderate	3,954 (2.8%)
	Severe	321 (0.2%)
	N/A or Not Documented	22,536 (16.1%)
	Missing	1,249 (0.9%)
Pulmonic Insufficiency	None	72,972 (52.3%)
	Trivial	20,655 (14.8%)
	Mild	6,531 (4.7%)
	Moderate	539 (0.4%)
	Severe	45 (0.0%)
	N/A or Not Documented	37,264 (26.7%)
	Missing	1,558 (1.1%)

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

We used the same dataset of isolated CABG operations from October 2014 to September 2015 for the entire report. The three exceptions are:

1. For validity testing and the comparison of participants over time, we used STS participants with procedures during both October 2013 - September 2014 and October 2014 - September 2015 time periods.
2. For the analysis of population disparities, current and over time, we used eligible patients from STS participants with procedures between October 2011 and September 2015 and defined relevant subgroups by age, gender, race, ethnicity and insurance status.
3. For the analysis on the impact of exclusions, we included the cases with contraindication for beta blockade at discharge.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We report trends of beta blockade at discharge among the following groups: Age (<75, ≥75), Gender, Race (White, Black and Other), Hispanic Ethnicity and Insurance (<65, ≥65).

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

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

Reliability is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. For this NQF submission, the measurement of interest is each participant's observed proportion. The true value is the proportion that would be observed hypothetically if the sample size was very large (i.e. infinite).

For the j-th participant, let n_j denote the number of eligible patients, let $y_j$denote the number of patients receiving beta-blockers, and let $\bar{y}_j = y_j/n_j$ denote the proportion of patients receiving beta-blockers. In addition, let μ_j denote the underlying true value of $\bar{y}_j$. To estimate reliability, we assumed the following hierarchical model for the data. At the first stage of the hierarchy, we assume that y_j is distributed according to a binomial distribution with sample size n_j and probability parameter μ_j. At the second stage of the hierarchy, we assumed that μ_j varies across participants according to a Beta distribution with mean $E[\mu_j] = \alpha/(\alpha+\beta)$ and $\text{var}[\mu_j] = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$, where α and β are unknown parameters to be estimated from the data. The unknown parameters α and β were estimated via maximum likelihood using the BETABIN macro for SAS software (BETABIN, version 2.2, 2005. Qi Statistics). The sample for this analysis included all **1,036 participants** and **139,564 eligible patients** in the main study period October 2014-September 2015. After estimating α and β, we then calculated the reliability that would be achieved if the measure were to be calculated on a sample size of 30 patients per participant. This estimated reliability was calculated as

$$\text{reliability} = [\text{corr}(\bar{y}, \mu)]^2 = \frac{1}{1 + (\hat{\alpha} + \hat{\beta})/n}$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote maximum likelihood estimates of α and β, respectively, and n =30. Because reliability increases with n, and because the vast majority of STS participants have >30 eligible patients per year, the reliability calculated with n =30 patients per participant provides a conservative lower bound for the actual reliability that will be achieved when the measure is applied to STS data from a 1 year period. Using the above formula, we also calculated the sample size n required per participant to achieve reliability of at least 0.50, 0.60, and 0.70, and the proportion of STS participants with at least this number of eligible patients in the most recent 1-year testing sample.

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

Estimated parameter values of the beta distribution were $\hat{\alpha}$=26.162 and $\hat{\beta}$= 0.5024. The estimated reliability with 30 eligible patients per participant was $1/(1 + (26.162 + 0.5024)/30) = 0.53$.

Based on these estimated parameter values, a sample size of 27 eligible patients per participant is needed to attain reliability of 0.50 and a sample size of 62 eligible patients per participant is needed to attain reliability of 0.70. During October 2014-September 2015, 95% of STS participants met the minimum required sample size for 0.50 reliability and 76% of STS participants met the minimum required sample size for 0.70 reliability.

	Reliability 0.50	**Reliability 0.60**	**Reliability 0.70**
Minimum required sample size per participant	27	40	62
Percent of participants meeting minimum sample size	95%	89%	76%

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

Reliability is comparable to or better than other NQF-endorsed STS outcome measures. The proposed measure has adequate statistical reliability to be used for confidential feedback reporting as well as public reporting.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Critical data elements

Participating sites are randomly selected for participation in STS Adult Cardiac Surgery Database Audit, which is designed to evaluate the accuracy, consistency, and comprehensiveness of data collection and ultimately validate the integrity of the data contained in the database. Telligen has conducted audits on behalf of STS since 2006. In 2015, 10% of STS Adult Cardiac Surgery Database participants (N=107) were audited. The audit process involves re-abstraction of data for 20 cases and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each of the 82 variables, each variable

category and overall. In 2015 the overall aggregate agreement rate was 96.17%, demonstrating that the data contained in the STS Adult Cardiac Surgery Database are both comprehensive and highly accurate.

Performance measure score

We calculated and compared the observed proportions of patients receiving the measure in the three performance groups. The measure has good face value if the three groups have different proportions as expected.

Face validity also implies that the measure is regarded as useful and valid by its intended users, including providers, consumers, payers, and regulators. The measure was developed with a panel of surgeon experts and statisticians. We have had near-universal acceptance of this measure by all stakeholders, with few if any relevant suggestions for change.

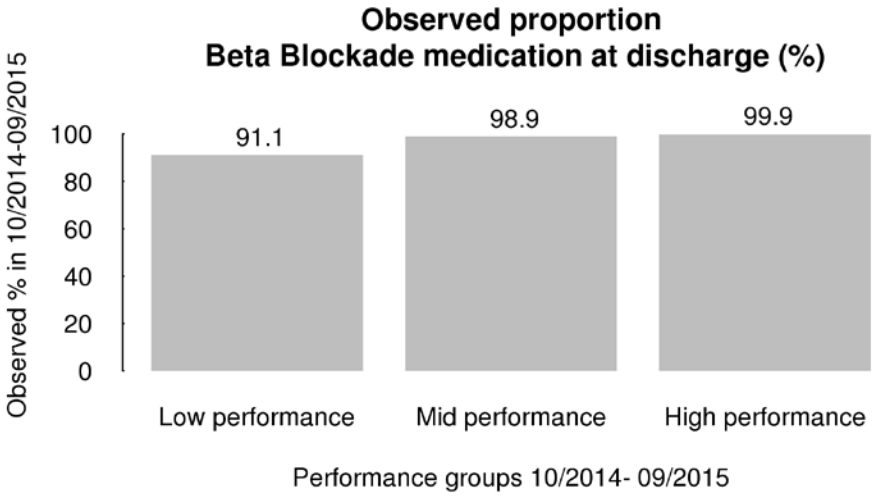
In addition, we tested the predictive validity of the measure. Predictive validity means that the results of this measure are predictive of future performance. We assessed the extent to which performance on this STS measure remains stable over time. In other words, does the measure at one point in time accurately predict performance at some later time?

The tests on validity used the concept of performance outliers to be more formally introduced in 2b5: Participants were labeled as "low performance" if the 95% exact binomial confidence interval of its event rate lies entirely below the population average (in other words, the upper bound of the 95% CI < population average). Participants were labeled as "high performance" if the 95% confidence interval lies entirely above 1. The remaining participants were labeled mid performance.

For each of the performance groups from the earlier period, we calculated the group specific measure proportions in the later period.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

STS participants deemed high performers by this measure have (on average) high rates of beta blockade at discharge. Thus, differences in performance were clinically meaningful as well as statistically significant. This is illustrated in the figure below using data from October 2014 to September 2015. Compared to participants who were deemed as having lower than average performance, those with better-than-average performance had higher rate of beta blockade at discharge (99.9% vs. 91.1%).



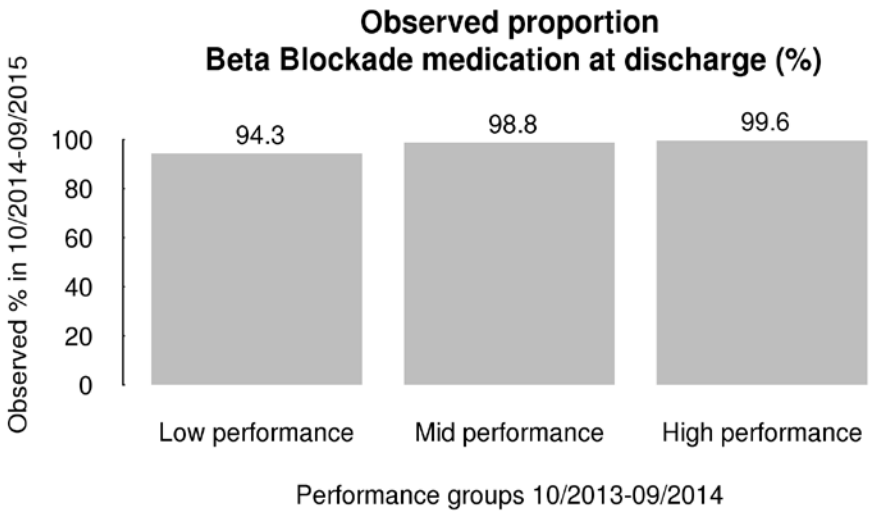
The predicted validity analysis was restricted to a sample of 1012 STS participants with patients receiving the measure in both time periods: October 2013 – September 2014 and October 2014 - September 2015. Among

participants who were high performance centers in October 2013 – September 2014, 76.1% of them were also high performers for October 2014 - September 2015. For comparison, only 5.2% of participants who were mid performers in October 2013 – September 2014 became high performers in October 2014 - September 2015. Thus, participants who performed better than average in October 2013 – September 2014 were over 14 times more likely to be identified as better performers in the next year. Similarly, participants who were low performance entities in the early year were more likely to remain low performers in the later year. Two participants jumped from low to high performing status (or vice versa) between the two adjacent 12-month periods. Thus, a consumer may reasonably expect that a high or low performer will likely be the same or became average in the near future, and a mid-performer is likely to remain average.

Change in performance categories between two time periods

		10/2014 – 09/ 2015		
		Low performance	Mid performance	High performance
10/2013 - 09/2014	Low performance	50	50	2
	Mid performance	39	780	45
	High performance	0	11	35

For each of the performance groups in the earlier period, we also calculated its aggregated proportion of patients receiving the measure in the later period. The aggregated proportions in the later periods were 99.6%, 98.8%, and 94.3% for the high, mid and low performance groups from the earlier period.



2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show that the measure reflects the proportion of patients who were discharged on beta blockers as designed, and that the past measure can be used to predict future performance. Together with face value, they support the validity of the measure.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded from the analysis cases if there was an in-hospital mortality or if beta blocker was contraindicated. We believe this is a clinically appropriate exclusion and is necessary to make the measure a consistent performance measure for the comparison across participants. The exclusion is precisely defined and specified.

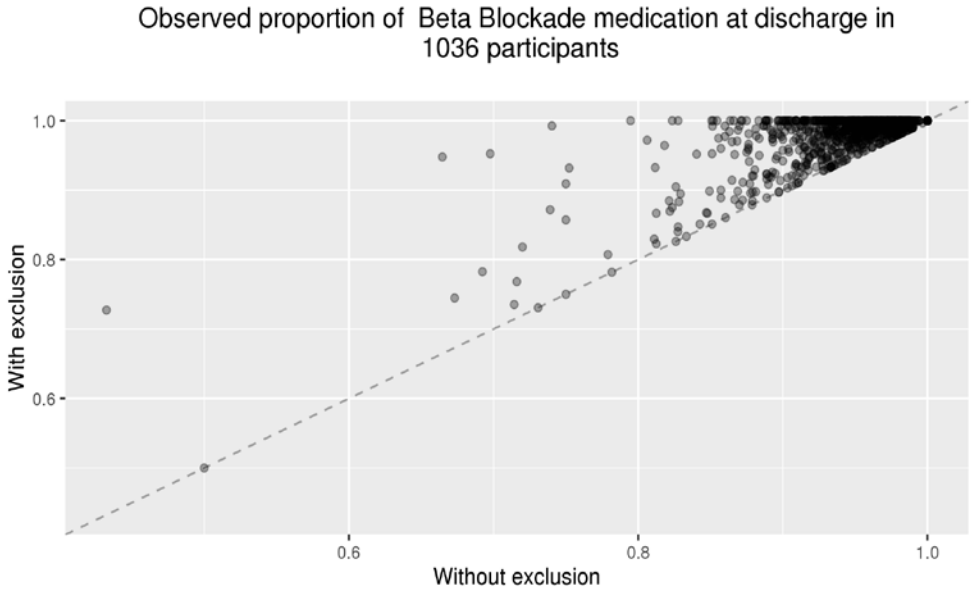
To show the impact of this exclusion, and how the measure would be distributed without it, we calculated and compared the distributions of the measure with and without the current exclusion criteria, with the exception of in-hospital deaths, that were excluded in all analyses.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1036	1036
# Operations	139564	144880
Mean	0.98	0.94
STD	0.039	0.051
IQR	0.024	0.046
0%	0.50	0.43
10%	0.94	0.89
20%	0.97	0.92
30%	0.98	0.94
40%	0.99	0.95
50%	0.99	0.96
60%	1.00	0.96
70%	1.00	0.97
80%	1.00	0.98
90%	1.00	0.99
100%	1.00	1.00
Midwest	296	296
Northeast	136	136
South	389	389
West	215	215
Low performance	94, 9.1%	103, 9.9%
Mid performance	859, 82.9%	830, 80.1%
High performance	83, 8.0%	103, 9.9%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.54. The Pearson correlation is 0.75.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

For the measure to consistently quantify the quality per its definition, it is necessary to exclude cases if there was an in-hospital mortality or if discharge beta blocker was contraindicated. It has an impact on the results for many participants, and the results would be distorted without these appropriate exclusions.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** 42T **risk categories**
☐ **Other,** 42T

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)*

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The summary statistic provided is the participant's observed proportion of eligible patients who receive beta blocker at discharge.

The degree of uncertainty surrounding an STS participant's beta blockade at discharge measure estimate is indicated by the 95% exact binomial confidence interval (CI) of its observed proportion. Point estimates and CI's of the observed proportion for an individual STS participant are reported along with a comparison to the STS average proportion of the study time period. A performance category interpretation is also given to STS participants. **Since higher value indicates better performance**, an STS participant is designated as having higher/lower than average performance for the measure if the 95% CI lies entirely **above/below** the STS average. The remaining participants are labeled as not distinguishable from the STS average performance. For the simplicity of this report, we call the three groups 'high performance', 'low performance' and 'mid performance', respectively.

The method is equivalent to performing an exact binomial test with the null hypothesis that the participant has the same proportion of patients receiving the measure as the population average. Those with a test p-value smaller than 0.05 are the low and high performance groups.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

As shown in the table below, the proportion of STS ACSD participants performing better and worse than STS average has remained similar over the last two 12-month periods. On average, more than 80% of the participants have performance indistinguishable from the STS average, and the remaining participants have performed differently.

Distribution	10/2013 – 09/2014 Observed Proportion	10/2014 – 09/2015 Observed Proportion
# Participant	1058	1036
# Operations	139921	139564
Low performance	109, 10.3%	94, 9.1%
Mid performance	902, 85.3%	859, 82.9%
High performance	47, 4.4%	83, 8.0%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The statistical test and the construction of confidence interval are widely used and accepted. The participants identified as having performed differently from the average likely have true performance characteristics that are different. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the amount of outliers the measure detects.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Due to great data quality, the source fields required by beta blockade at discharge had only 0.3% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with greater than 5% missing data are excluded from the calculation of the measure.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Overall 0.3% of data were missing. 99% of participants had missing rate of 4% or lower. Ten out of 1048 participants were not included because of having missing rates higher than 5%.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The rates of missing data in the STS Adult Cardiac Surgery Database were very low and are getting lower. We therefore concluded that systematic missing data did not lead to bias in our measure.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Data Collection:
There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Public Reporting This measure is one of eleven component measures of the STS CABG Composite Score. Approximately 49.8% of STS Adult Cardiac Surgery Database participants are voluntarily enrolled in the STS public reporting program. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online and Consumer Reports Health: www.ConsumerReports.org/hospitalratings Quality Improvement with Benchmarking (external benchmarking to multiple organizations) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database Quality Improvement (Internal to the specific organization) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database

4a.1. **For each CURRENT use, checked above, provide:**

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

Please see table above

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

N/A

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

N/A

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Please see sections 1b.2 and 1b.4

In the table below we provide the overall trend over time of the measure performance at the patient level. The aggregate proportion of eligible patients receiving beta blockade at discharge was computed for each time period. Although measure performance was

above 97.5% during the entire period, we can see a moderate increase on the proportion of patients receiving beta blockade over the last 4 years.

	10/2011 - 09/2012		10/2012 - 09/2013		10/2013 – 09/2014		10/2014-09/2015	
All	97.96%	98.47%	98.68%	98.53%				

Geographic area and number and percentage of accountable entities and patients included
Number of participants and operations by geographic regions, during the two last consecutive time periods, October 2013-September 2014 and October 2014-September 2015.

	10/2013 – 09/2014				10/2014 – 09/2015			
	Midwest	Northeast	South	West	Midwest	Northeast	South	West
# Participant	306	134	398	220	296	136	389	215
% Participant	28.9%	12.7%	37.6%	20.8%	28.6%	13.1%	37.5%	20.8%
# Operation	34439	22272	61365	21845	34154	22351	60956	22103
% Operation	24.6%	15.9%	43.9%	15.6%	24.5%	16.0%	43.7%	15.8%

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2015, 10% of participants, i.e., 107 facilities, were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0116 : Anti-Platelet Medication at Discharge
0118 : Anti-Lipid Treatment Discharge
0119 : Risk-Adjusted Operative Mortality for CABG
0127 : Preoperative Beta Blockade
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection

0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
0134 : Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** 0117_Beta_Blockade_at_Discharge_Appendix_-_S.9-_1b.2-_1b.4-_Guidelines.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
- Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery

- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2004

Ad.3 **Month and Year of most recent revision:** 06, 2016

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

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0134
De.2. **Measure Title:** Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
Co.1.1. **Measure Steward:** The Society of Thoracic Surgeons
De.3. **Brief Description of Measure:** Percentage of patients aged 18 years and older undergoing isolated coronary artery bypass graft (CABG) who received an internal mammary artery (IMA) graft
1b.1. **Developer Rationale:** Use of the internal mammary artery as coronary bypass conduit has definitively and repeatedly been shown to substantially increase patient survival in the long term. Using this measure should encourage, and potentially increase, the use of the internal mammary arteries as coronary bypass conduits.

S.4. **Numerator Statement:** Number of patients undergoing isolated coronary artery bypass graft (CABG) who received an internal mammary artery (IMA) graft
S.7. **Denominator Statement:** Patients undergoing isolated CABG
S.10. **Denominator Exclusions:** Cases are removed from the denominator if the patient had a previous CABG prior to the current admission or if IMA was not used and one of the following reasons was provided:

- Subclavian stenosis
- Previous cardiac or thoracic surgery
- Previous mediastinal radiation
- Emergent or salvage procedure
- No (bypassable) LAD disease

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: May 09, 2007 **Most Recent Endorsement Date:** Jan 31, 2012

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

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

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a

systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** **X Yes** **☐ No**
- **Quality, Quantity and Consistency of evidence provided?** **X Yes** **☐ No**
- **Evidence graded?** **X Yes** **☐ No**

Evidence Summary with **Summary of prior review**

- During previous review, the Committee agreed there was strong evidence to support the measure.
- The 2011 ACCF/AHA Guideline for Coronary Artery Bypass Graft Surgery includes recommendations for use of internal mammary arteries as follows:
  - If possible, the left internal mammary artery (LIMA) should be used to bypass the left anterior descending (LAD) artery when bypass of the LAD artery is indicated. *(Class I, Level of Evidence: B)*;
  - The right internal mammary artery is probably indicated to bypass the LAD artery when the LIMA is unavailable or unsuitable as a bypass conduit. *(Class II, Level of Evidence: C);*
  - When anatomically and clinically suitable, use of a second IMA to graft the left circumflex or right coronary artery (when critically stenosed and perfusing LV myocardium) is reasonable to improve the likelihood of survival and to decrease reintervention. *(Class II, Level of Evidence: B)*
- Evidence submitted at this and the last review included observational, retrospective, prospective studies – randomized controlled trials that demonstrated the value of using the IMA in coronary artery bypass graft surgery.

Changes to evidence from last review
The developer provided updated evidence for this measure:

- Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.
- Raza S, Sabik JF 3rd, Masabni K, Ainkaran P, Lytle BW, Blackstone EH. Surgical revascularization techniques that minimize surgical risk and maximize late survival after coronary artery bypass grafting in patients with diabetes mellitus. J Thorac Cardiovasc Surg. 2014. Oct;148(4):1257-1264; discussion 1264-6.
- Locker C, Schaff HV, Dearani JA, Joyce LD, Park SJ, Burkhart HM, et al. Multiple arterial grafts improve late survival of patients undergoing coronary artery bypass graft surgery: analysis of 8622 patients with multivessel disease. Circulation 2012;126:1023-30.

The evidence includes the guideline cited above and is directionally the same compared to that for the previous NQF review.

Questions for the Committee:

- *Does the Committee agree that the evidence continues to support use of the IMA graft in patients undergoing CABG as specified in the measure?*
- *Does the Committee believe there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm: Process measure/systematic review and grading evidence (Box 3) → Information on QQC presented (Box 4) → SR conclusion (5b) → Moderate

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

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

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

- With a performance compliance mean of 94.7% and median of 95.8%, the Committee saw opportunity for improvement and noted that the literature reported that IMA was used less often in women.

- Most recent data are drawn from STS database participant-specific proportions of patients with IMA use in CABG during the period from October 2014 – September 2015. Eligible operations are included with patient-specific detail provided regarding age, sex, race, ethnicity, insurance, bsa, acuity and more than 20 clinical conditions.
- Performance ranged from 71.8% to 100% with mean performance during this period at 98.5% (1,041 STS database participants; 134,689 operations). Geographic distribution of participants is presented. Sample size, data and other factors provide detail to appreciate the gap.
- Of note, per the developer, more than 90% of cardiothoracic programs in the US are STS database participants.

Disparities

- The developer reports that for analysis of disparities, eligible patients from STS database participants with procedures between 2011 and September 2015 were used. Relevant subgroups were defined by age, gender, race, and insurance status.
- The performance ranges from 2011-2012 to 2014-2015 are presented below. In each group, performance improved or stayed at a relatively steady state from year to year. Year of lowest and highest rates are included if they are not the earliest (low) or latest (high).
  - Gender – Male, 98.65% (year ending 9/2013) – 99.05%; Female, 97.49% - 98.07%
  - Age Groups - <75, 98.58% - 98.98%; >=75, 97.39% (year ending 9/2013) – 98.11%
  - Race – White, 98.45% - 98.91%; Black, 97.67% (year ending 9/2013) – 98.17% (year ending 9/2014; Other, 97.84% (year ending 9/2013) – 98.60%
  - Insurance - Age>=65, lowest performance 97.08% (Medicare+Medicaid – 9/2013) to 98.75% (Medicare w/o Medicaid/Commercial) as highest in 2014 - 2015 period and for those in the Age<65, the low in 2012 - 2013 period of 98.20% (Medicare/Medicaid) to high in 2013-2014 of 98.78% (Medicare/Medicaid). The data suggests relatively uniform high use of IMA across all groups.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *How should the disparities information be factored into consideration of sociodemographic factors going forward?*
- *Assuming the Committee continues to view this measure as highly credible, reliable and valid, should the measure be considered for Inactive Endorsement with Reserve Status on the basis of performance gap?*

Preliminary rating for opportunity for improvement: ☐ High ☐ Moderate X Low ☐ Insufficient

Committee pre-evaluation comments

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

1a.

- Very relevant process measure strongly supported by evidence
- Process measure, but with strong linkage
- Highly actionable at the single surgeon level
- Part of the composite CABG outcome
- Topped out?
- The evidence supporting this measure has been extensively reviewed in previous meetings. I do not see any new information that would change the prior conclusions. Use of IMA in CABG continues to be considered standard of care. I do not believe further discussion on this issue is necessary and support the staff recommendation of MODERATE.

1b.

- Unclear.
  - Is there a bell curve displaying prevalence of low performers?
  - 2b2: 'IMA use varied from 96.2% to 99.6%'.
  - also p 31-32 for graph-lo/mid/hi performance groups migration over time = performance gap?
- Performance is uniformly high across STS participating centers. The appendix suggests that even institutions performing at the 20 %ile are 98% effective. The reported data reflect a very slight improvement in the past year. I agree with NQF staff that the performance gap is LOW. I would suggest the committee consider reserve status for this measure if this would not negatively impact its use in the CABG composite measure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The measure is specified for analysis at the group/practice and facility levels of analysis and intended for use in the hospital/acute care setting.
- The measure assesses the number of patients undergoing CABG who received an internal mammary artery (IM graft). Denominator exclusions are subclavian stenosis, previous cardiac or thoracic surgery, previous mediastinal radiation, emergent or salvage procedure, no (bypassable) LAD disease. (*Of note, the developer removed IMA suitability as an exclusion, as recommended by the Surgery Committee during the previous evaluation*.)
- The developer notes there have been no changes to the measure specifications since it was last endorsed. As noted above IMA suitability was removed as recommended by the Surgery Committee.
- The data source for the measure is the STS Adult Cardiac Surgery Database. Data is collected using the STS database collection form (version 2.81) that includes detailed items regarding a wide range of factors including use of IMA (left, right, both, no, and if no, primary reason including the exclusions.
- The measure is not risk adjusted or risk stratified.

Questions for the Committee :

- *Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Previous reliability testing included inter-rater reliability testing of 40 randomly selected sites participating in the STS Adult Cardiac Surgery Database.

SUMMARY OF TESTING

Reliability testing level **X Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure **X Yes** ☐ **No**

Method and Results of reliability testing

- Testing was done using the sample of 1,041 STS participants (134,689 operations) who submitted data between October 2014 and September 2015. Exclusions are cases with specifically defined reasons for not having IMA used in CABG or with previous CABG.
- Score level testing was done using a method described as an equivalent to signal-to-noise analysis. The developer notes that reliability increases with number of patients and the vast majority of STS participants have >30 eligible patients per year thus calculating reliability with 30 patients per participant provides a conservative lower bound for the actual reliability that will be achieved when the measure is applied to STS data from a 1 year period.
- The developer states that the estimated reliability with 30 eligible patients per participant = 0.36. The table below provides information about the minimum sample sizes to achieve for reliability at 3 additional levels along with the percent of STS database participants who meet the minimum sample size.

	Reliability 0.50	Reliability 0.60	Reliability 0.70
Minimum required sample size per participant	54	81	126
Percent of participants meeting minimum sample size	80%	65%	41%

- The developer interprets the results as adequate statistical reliability for use in confidential feedback and public reporting.

Questions for the Committee:

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

Guidance from the Reliability Algorithm: Precise specifications (Box 1) → Empiric reliability testing (Box 2) →Testing at measure score level (Box 4) → Method described and appropriate (Box 5) → Level of confidence (Box 6)

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. X Yes ☐ Somewhat ☐ No

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- At prior maintenance submission, validity testing was conducted using the STS database audit process that involved 40 randomly selected sites (of the 628 eligible database participants), designated for audit to evaluate accuracy, consistency and comprehensiveness of data collection and integrity.

SUMMARY OF TESTING

Validity testing level ☐ Measure score ☐ Data element testing against a gold standard X Both

Method of validity testing of the measure score:

☐ **Face validity only**
X Empirical validity testing of the measure score

Validity testing method:

- Data element testing was done using the STS Adult Cardiac Database Audit conducted by a third party, by which participant sites are randomly selected, on an annual basis, to undergo audit of the accuracy, consistency, and comprehensiveness of data collection. In 2015, 10 percent of STS Adult Cardiac Surgery Database participants (107) were audited. Auditing involved re-abstraction of data for 20 cases from each audited participant and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each variable, each variable category and overall. Overall agreement was 96.17%.
- For validity testing and comparison of participants over time, STS participants with procedures during both October 2013 – September 2014 and October 2014 – September 2015 were used.
- Performance measure score testing was done using face validity and predictive validity - assessing stability of

performance over time.

- Predictive validity was assessed using a concept of "outliers". The developers posit that stability of measure scores over time may indicate that the measure is capturing an accurate indication of provider performance. ***There is some disagreement about whether stability in performance demonstrates predictive validity; some would argue that changes in performance over time are to be expected—and are, in fact, desirable—as the result of quality improvement interventions. NQF guidance suggests that predictive validity should compare measure results to another measure of the same construct or to a different outcome measure.***
- Participants were placed into three groups – low performance (95% exact binomial confidence interval of event rate entirely below population average), high performance (95% confidence interval entirely above 1) and mid performance (remaining participants). Predictive validity analysis was restricted to a sample of 1,013 participants that received the measure in both time periods (10/2013 – 9/2014 and 10/2014 – 9/2015).
- Of high performers in the earlier period, 21.1% were high performers in the second period. Only 1.6% of mid performers in the early period became high performers in the second period. Participants that performed better than average in the earlier period were over 12 times more likely to be better performers in the next year and those identified as low performers were likely to remain so in the next year.
- The aggregated proportion of patients for the low, mid and high performance groups from the first time period that received IMA in the later period (10/2014 – 9/2015) was calculated and is reported as 96.2% (low performance), 98.9% (mid performance), and 99.6% (high performance); i.e., IMA use varied from 96.2% to 99.6%. This information demonstrates that performance remained constant and high over the period described. The developer concludes that the results of the measure can be used to predict future performance.

Questions for the Committee:

- *Does the testing that is described validate the measure and allow conclusions about quality?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*
- *What does the score suggest in terms of opportunity for improvement?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- To assess impact of the measure exclusions, the distribution of the measure scores with and without exclusion was computed.
- The exclusions to the measure are a previous CABG prior to the current admission or if IMA was not used and one of the following reasons was provided: subclavian stenosis**,** previous cardiac or thoracic surgery**,** previous mediastinal radiation**,** emergent or salvage procedure, and no (bypassable) LAD disease.
- The developer has calculated the distribution of the participant-specific observed proportion of patients receiving IMA with and without the exclusion for the 10/2014 – 9/2015 time period and shown the percent change across each of the three performance groups to demonstrate its effect on measure results.
- With respect to the exclusion criteria , proportion of patients receiving IMA changed as follows:
  - low performance participants with exclusion, 7.3%; without,10.7%;
  - mid performance 90.7% with exclusion; without, 77.8%;
  - high performance 2.0%; without, 11.5%.
- The developer reports that STS database participants performing better and worse than the STS average has remained similar over the two time periods, 10/2013 – 9/2014 and 10/2014 – 9/2015, with more than 90% having performance indistinguishable from the STS average.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *What do the differences tell the Committee about justification for the exclusions?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** **X None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

As noted in earlier sections, the developer has presented information about low, mid, and high performance participants.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed. The measure uses a single data source and has one set of specifications.

2b7. Missing Data

Overall rate of missing data is reported as 0.1%. Missing data are imputed to "no" (use of IMA); participants with greater than 5% (7 of 1,048 participants) missing data are excluded from the measure calculation. The developer reported that 99% of participants had a missing rate of 4% or lower.

Guidance from the Validity Algorithm Specifications consistent with evidence (Box 1) → Threats to validity (Box 2) → Empirical validity testing (Box 3) → Validity systematically assessed (Box 4) → Confidence that scores are indicator of quality (Box 5)

Preliminary rating for validity: ☐ High X Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

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

2a1.

- No concerns
- The specifications have not changed since initial endorsement. The data can be extracted from STS database participants in a highly relievable manner. The staff asks whether data can be reliably abstracted by non-STS participants. In my opinion that data should be able to be readily abstracted by any cardiac surgical center. Further, there are very few centers in the US that do not participate in the STS database. I have no concerns here.

2b1.

- specs consistent w evidence
- The specifications seem consistent with the evidence. The staff asks whether the degree of differences seem because of the exclusions justifies the collection "burden." In this reviewers opinion, the exclusions are justified and essential to the face validity o the measure.

2b2.

- reliable
- I believe reliability testing was adequate and unchanged and agree with the staff assessment of LOW. The reliability is less than optimal but has not changed from prior reviews of this measure.

2b3.

- no
- I have no significant additions to the staff comments here.

Criterion 3. Feasibility

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

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

The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.

Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants. Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.

There are no additional costs for data collection specific to the measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement. STS analyses indicate that the STS database includes more than 90% of cardiothoracic programs in the US.

Questions for the Committee:
Is the effort and cost associated with abstracting the required data elements appropriate to the value of the measure?

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

Committee pre-evaluation comments

Criteria 3: Feasibility

- 90% USA cardiac surgery programs participle in STS DB; is feasible
- I see no meaningful changes from previous assessments of feasibility and agree with the staff assessment of MODERATE. The staff asks if the effort in collection is justified by the information obtained. This is relevant to whether reserve status is appropriate and is discussed elsewhere.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

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

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

Accountability program details

- STS analyses indicate that the STS database includes more than 90% of cardiothoracic programs in the US.
- The measure is:
  - 1 of 11 component measures of the STS CABG Composite Score reported by STS Public Reporting Online and Consumer Reports Health (About 49.8% of the STS Adult Cardiac Surgery Database's 1,100 participants are voluntarily enrolled in the public reporting program.);
  - reported to CMS on behalf of consenting surgeons (Physician Quality Reporting System measure #43;

- used for QI with benchmarking (involves external benchmarking with multiple organizations); and used for
- QI internal to the individual organization.

Improvement results:

- Aggregate proportion of eligible patients receiving IMA above 98% with range from a low of 98.36% for the period 10/2011 – 9/2012 to a high of 98.81% in 10/2014 – 9-2015.

Unexpected findings (positive or negative) during implementation : None identified

Potential harms

- The developer reports it is not aware of any negative unintended consequences, noting that all public reporting has potential for such things as gaming and risk aversion. The developer attempts to control gaming though its audit process and risk aversion by accounting for the expected risk for providers who care for sicker patients.

Feedback :

- Per the June 2012 Surgery Endorsement Maintenance report, public comment noted that the measure should retain endorsement and be placed in reserve status.

***Questions for the Committee**:*

o *Does the measure continue to be useful for improvement given the high level of performance?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- publically reported
- This is a widely used and accepted measure. My only concern here is whether this measure is close to being "topped out." I favor continued endorsement of the measure in reserve status as above.

Criterion 5: Related and Competing Measures

Related or competing measures

- The developer lists the following related or competing measures:

0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0116 : Anti-Platelet Medication at Discharge
0117 : Beta Blockade at Discharge
0118 : Anti-Lipid Treatment Discharge
0119 : Risk-Adjusted Operative Mortality for CABG
0127 : Preoperative Beta Blockade
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

Harmonization

- The developer reports the measure specifications are completely harmonized.

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0134

Measure Title: Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 0696 STS CABG Composite Score

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☐ Health outcome:

☐Patient-reported outcome (PRO): 42T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 42T

☒ Process: Use of IMA in CABG

☐ Structure: 42T

☐ Other: 42T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

The superiority of internal mammary arteries over saphenous vein grafts as coronary artery bypass conduits has been known for at least 25 years. The overwhelming evidence came initially both from retrospective reviews and randomized controlled trials. The Cleveland Clinic showed in a 10 year review in 1986 that survival after coronary bypass grafting was improved if an internal mammary artery was placed to the left anterior descending

coronary artery versus a saphenous vein graft. A randomized controlled trial, begun in 1975, with 10 year follow-up on 80 patients gave similar results. Since then, a plethora of studies, including The Society of Thoracic Surgeons Adult Cardiac database evaluation, have continued to prove that patients with internal mammary artery grafts, especially to the left anterior descending coronary artery, live longer than any other conduit combination. Most, if not all, of this benefit is derived from the improved long-term patency rates associated with internal mammary arteries over other conduits. This observation is also well documented in the literature.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.

http://circ.ahajournals.org/content/124/23/e652

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page e132

2.1.4. Bypass Graft Conduit: Recommendations

Class I Recommendation

If possible, the left internal mammary artery (LIMA) should be used to bypass the left anterior descending (LAD) artery when bypass of the LAD artery is indicated. (Level of Evidence: B)

Class IIa Recommendation

1. The right internal mammary artery (IMA) is probably indicated to bypass the LAD artery when the LIMA is unavailable or unsuitable as a bypass conduit. (Level of Evidence: C)

2. When anatomically and clinically suitable, use of a second IMA to graft the left circumflex or right coronary artery (when critically stenosed and perfusing LV myocardium) is reasonable to improve the likelihood of survival and to decrease reintervention. (Level of Evidence: B)

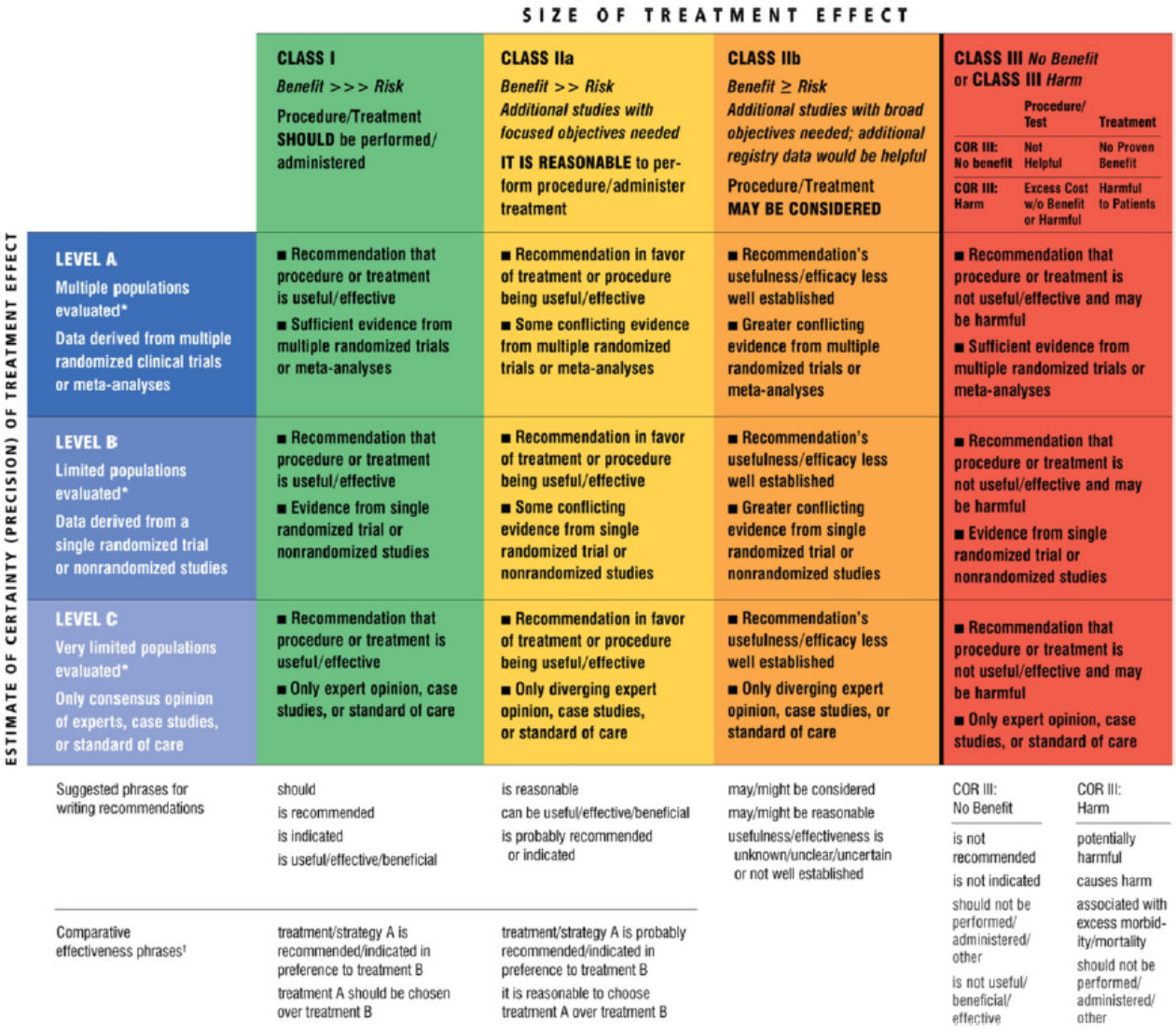
1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Class 1 Level B. Recommendation that procedure or treatment is useful/effective. Evidence from single randomized trial or nonrandomized studies

Class IIa Level B. Recommendation in favor of treatment or procedure being useful/effective. Some conflicting evidence from single randomized trial or nonrandomized studies

Class IIa Level C. Recommendation in favor of treatment or procedure being useful/effective. Only diverging expert opinion, case studies or standard of care

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)



SIZE OF TREATMENT EFFECT

ESTIMATE OF CERTAINTY (PRECISION) OF TREATMENT EFFECT	CLASS I *Benefit >>> Risk* Procedure/Treatment **SHOULD** be performed/ administered	CLASS IIa *Benefit >> Risk* *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	CLASS IIb *Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	CLASS III *No Benefit* or CLASS III *Harm* Procedure/Test / Treatment COR III: No benefit — Not Helpful / No Proven Benefit COR III: Harm — Excess Cost w/o Benefit or Harmful / Harmful to Patients
LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical trials or meta-analyses	■ Recommendation that procedure or treatment is useful/effective ■ Sufficient evidence from multiple randomized trials or meta-analyses	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Sufficient evidence from multiple randomized trials or meta-analyses
LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is useful/effective ■ Evidence from single randomized trial or nonrandomized studies	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Evidence from single randomized trial or nonrandomized studies
LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	■ Recommendation that procedure or treatment is useful/effective ■ Only expert opinion, case studies, or standard of care	■ Recommendation in favor of treatment or procedure being useful/effective ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation's usefulness/efficacy less well established ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Only expert opinion, case studies, or standard of care
Suggested phrases for writing recommendations	should is recommended is indicated is useful/effective/beneficial	is reasonable can be useful/effective/beneficial is probably recommended or indicated	may/might be considered may/might be reasonable usefulness/effectiveness is unknown/unclear/uncertain or not well established	COR III: No Benefit: is not recommended; is not indicated; should not be performed/ administered/ other; is not useful/ beneficial/ effective COR III: Harm: potentially harmful; causes harm; associated with excess morbidity/mortality; should not be performed/ administered/ other
Comparative effectiveness phrases†	treatment/strategy A is recommended/indicated in preference to treatment B treatment A should be chosen over treatment B	treatment/strategy A is probably recommended/indicated in preference to treatment B it is reasonable to choose treatment A over treatment B		

A recommendation with Level of Evidence B or C does not imply that the recommendation is weak. Many important clinical questions addressed in the guidelines do not lend themselves to clinical trials. Although randomized trials are unavailable, there may be a very clear clinical consensus that a particular test or therapy is useful or effective.

*Data available from clinical trials or registries about the usefulness/efficacy in different subpopulations, such as sex, age, history of diabetes, history of prior myocardial infarction, history of heart failure, and prior aspirin use. †For comparative effectiveness recommendations (Class I and IIa; Level of Evidence A and B only), studies that support the use of comparator verbs should involve direct comparisons of the treatments or strategies being evaluated.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

ACCF/AHA Task Force on Practice Guidelines. Methodologies and Policies from the ACCF/AHA Task Force on Practice Guideline. June 2010.

http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf and

http://circ.ahajournals.org/site/manual/index.xhtml

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Information relevant to this section is provided in detail in the previous sections and referenced guideline. Please see Appendix.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Survival, postoperative myocardial infarction, hospitalization for cardiac events, need for reoperation, and recurrence of angina.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

See 1a.7

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.7

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 42T

See 1a.7

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

See 1a.7

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

See 1a.7

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

See 1a.7

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

See 1a.7

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

See 1a.7

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1a._Evidence_-_0134_Use_of_IMA_in_CABG.docx

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Use of the internal mammary artery as coronary bypass conduit has definitively and repeatedly been shown to substantially increase patient survival in the long term. Using this measure should encourage, and potentially increase, the use of the internal mammary arteries as coronary bypass conduits.

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

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

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from**

the literature that addresses disparities in care on the specific focus of measurement. Include citations.
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
The internal mammary artery has definitively and repeatedly been shown to be the best conduit for coronary bypass grafting. It has been shown to have the highest patency rates compared to other conduits and its use substantially increases patient survival in the long term over other conduit choices.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
- Raza S, Sabik JF 3rd, Masabni K, Ainkaran P, Lytle BW, Blackstone EH. Surgical revascularization techniques that minimize surgical risk and maximize late survival after coronary artery bypass grafting in patients with diabetes mellitus. J Thorac Cardiovasc Surg. 2014. Oct;148(4):1257-1264; discussion 1264-6.
- Locker C, Schaff HV, Dearani JA, Joyce LD, Park SJ, Burkhart HM, et al. Multiple arterial grafts improve late survival of patients undergoing coronary artery bypass graft surgery: analysis of 8622 patients with multivessel disease. Circulation 2012;126:1023-30.
- Ferguson TB Jr, Coombs LP, Peterson ED. Internal thoracic artery grafting in the elderly patient undergoing coronary artery bypass grafting: room for process improvement? J Thorac Cardiovasc Surg. 2002;123(5):869-880.
- Leavitt B, O'Connor GT, et al. Use of the internal mammary artery graft and in-hospital mortality and other adverse outcomes associated with coronary artery bypass surgery. Circulation. 2001;103(4):507-512.
- Loop FD, Lytle BW, Cosgrove DM, et al. Influence of the internal-mammary-artery graft on 10-year survival and other cardiac events. N Engl J Med. 1986 Jan 2;314(1):1-6.
- Lytle BW, Blackstone EH, Loop FD, et a. Two internal thoracic artery grafts are better than one. J Thorac Cardiovasc Surg. 1999 May;117(5):855-72.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;
http://www.sts.org/sites/default/files/documents/STSAdultCVDataSpecificationsV2_81.pdf

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
None

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients undergoing isolated coronary artery bypass graft (CABG) who received an internal mammary artery (IMA) graft

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Denominator – 12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of isolated CABG procedures in which IMA Artery Used [IMAArtUs (STS Adult Cardiac Surgery Database Version 2.81] is marked "Left IMA," "Right IMA," or "Both IMAs"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients undergoing isolated CABG

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Number of isolated CABG procedures excluding cases that were a previous CABG prior to the current admission or if IMA was not used and one of the acceptable reasons was provided. The SQL code used to create the function used to identify cardiac procedures is provided in the Appendix.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Cases are removed from the denominator if the patient had a previous CABG prior to the current admission or if IMA was not used and one of the following reasons was provided:
- Subclavian stenosis
- Previous cardiac or thoracic surgery

- Previous mediastinal radiation
- Emergent or salvage procedure
- No (bypassable) LAD disease

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Patients with previous CABG, identified where PrCAB is marked "yes"

or

IMA Artery Used (IMAArtUs) is marked “no IMA” and primary reason for no IMA (NoIMARsn) is marked as any of the following:
- Subclavian stenosis
- Previous cardiac or thoracic surgery
- Previous mediastinal radiation
- Emergent or salvage procedure
- No (bypassable) LAD disease

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
The source fields required by the IMA use in CABG measure had only 0.1% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with greater than 5% missing data are excluded from the calculation of the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database Version 2.81

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
2.1_Testing_-_0134_Use_of_IMA_in_CABG.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0134
Measure Title: Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
Date of Submission: 6/5/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion

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

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 42T	☐ other: 42T

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

STS Adult Cardiac Surgery Database (ACSD) Version 2.81

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

October 2014 – September 2015

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 42T	☐ other: 42T

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

The calculation of the IMA use in CABG measure of the 12 months from October 2014 to September 2015 used 134,689 operations from 1,041 STS participants.

Distribution of participant sample sizes (denominator), and observed proportion of patients receiving the measure (numerator/denominator)

Stat	N	% IMA use
N	1041.0	1041.0
Mean	135.8	98.5
STD	107.7	2.6
IQR	113.0	1.9
0%	1.0	71.8
10%	38.0	96.2
20%	55.0	97.7
30%	72.0	98.4
40%	87.0	98.9
50%	105.0	99.4
60%	128.0	100.0
70%	158.0	100.0
80%	201.0	100.0
90%	276.0	100.0
100%	869.0	100.0

Distribution of participants by geographic regions

REGION	
Midwest	297
Northeast	136
South	393
West	215

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

All eligible isolated operations were included except cases with reasons for not having IMA in CABG or with previous CABG.

	Effects	Overall N=141347
Age (years)	Median (IQR)	66.0 (58.0, 72.0)
	Missing	0 (0.0%)
Sex	Male	106,413 (75.3%)
	Female	34,872 (24.7%)
	Missing	62 (0.0%)
Race - Asian	No	133,795 (94.7%)
	Yes	4,451 (3.1%)
	Missing	3,101 (2.2%)

	Effects	Overall N=141347
Race - Black / African American	No	127,606 (90.3%)
	Yes	10,641 (7.5%)
	Missing	3,100 (2.2%)
Race - White	No	21,243 (15.0%)
	Yes	117,068 (82.8%)
	Missing	3,036 (2.1%)
Race - American Indian / Alaskan Native	No	137,332 (97.2%)
	Yes	916 (0.6%)
	Missing	3,099 (2.2%)
Race - Other	No	133,311 (94.3%)
	Yes	4,568 (3.2%)
	Missing	3,468 (2.5%)
Native Hawaiian / Pacific Islander	No	137,518 (97.3%)
	Yes	669 (0.5%)
	Missing	3,160 (2.2%)
Hispanic or Latino Ethnicity	No	123,891 (87.7%)
	Yes	10,054 (7.1%)
	Missing	7,402 (5.2%)
Insurance: Younger than 65	Medicare/Medicaid	17,585 (27.2%)
	Commercial/HMO	38,540 (59.6%)
	None/Self Paid	5,158 (8.0%)
	Other	3,328 (5.2%)
Insurance: 65 or Older	Medicare+Medicaid	4,856 (6.3%)
	Medicare+Commercial without Medicaid	42,223 (55.0%)
	Medicare without Medicaid/Commercial	29,657 (38.6%)
Region	NORTHEAST	22,655 (16.0%)
	SOUTH	61,382 (43.4%)
	MIDWEST	34,528 (24.4%)
	WEST	22,782 (16.1%)
Body Surface Area (m)	<1.5	1,863 (1.3%)
	>=1.5 and <1.75	17,038 (12.1%)
	>=1.75 and <2	48,580 (34.4%)
	>=2	73,757 (52.2%)
	Missing	109 (0.1%)
Diabetes	No Diabetes	73,062 (51.7%)
	Diabetes - Noninsulin	41,632 (29.5%)
	Diabetes - Insulin	25,112 (17.8%)
	Diabetes - Other	371 (0.3%)
	Diabetes - Missing Treatment	807 (0.6%)
	Missing	363 (0.3%)
Hypertension	No	15,857 (11.2%)

	Effects	Overall N=141347
	Yes	125,139 (88.5%)
	Missing	351 (0.2%)
Renal Function	Creatinine <1 mg/dL	68,120 (48.2%)
	Creatinine 1-1.5 mg/dL	56,880 (40.2%)
	Creatinine 1.5-2 mg/dL	8,434 (6.0%)
	Creatinine 2-2.5 mg/dL	1,823 (1.3%)
	Creatinine >2.5 mg/dL	1,439 (1.0%)
	Dialysis	4,314 (3.1%)
	Missing	337 (0.2%)
Dyslipidemia	No	17,241 (12.2%)
	Yes	123,437 (87.3%)
	Missing	669 (0.5%)
Chronic Lung Disease (CLD)	None	101,835 (72.0%)
	Mild	14,966 (10.6%)
	Moderate	6,875 (4.9%)
	Severe	6,063 (4.3%)
	5	6,932 (4.9%)
	Missing	4,676 (3.3%)
Peripheral Vascular Disease (PVD)	No	120,679 (85.4%)
	Yes	19,699 (13.9%)
	Missing	969 (0.7%)
Cerebrovascular Disease (CVD)	No CVD	113,040 (80.0%)
	CVD-NO CVA	28,307 (20.0%)
Endocarditis	No Endocarditis	141,052 (99.8%)
	Treated Endocarditis	67 (0.0%)
	Active Endocarditis	8 (0.0%)
	Endocarditis - Missing Type	7 (0.0%)
	Missing	213 (0.2%)
Acuity Status	Elective	53,012 (37.5%)
	Urgent	82,561 (58.4%)
	Emergent	5,637 (4.0%)
	Emergent Salvage	116 (0.1%)
	Missing	21 (0.0%)
Myocardial Infarction	No Prior MI	66,297 (46.9%)
	MI >21 days	25,864 (18.3%)
	MI 8-21 days	6,883 (4.9%)
	MI 1-7 days	35,809 (25.3%)
	MI 6-24 hrs	3,381 (2.4%)
	MI <= 6 hrs	1,592 (1.1%)
	MI - Missing Timing	355 (0.3%)
	Missing	1,166 (0.8%)
Cardiogenic Shock	No	139,517 (98.7%)

	Effects	Overall N=141347
	Yes	1,730 (1.2%)
	Missing	100 (0.1%)
Preop IABP	No	131,059 (92.7%)
	Yes	10,096 (7.1%)
	Missing	192 (0.1%)
Congestive Heart Failure	No CHF	112,614 (79.7%)
	CHF NYHA-I	2,344 (1.7%)
	CHF NYHA-II	8,189 (5.8%)
	CHF NYHA-III	9,961 (7.0%)
	CHF NYHA-IV	5,977 (4.2%)
	CHF Missing NYHA	977 (0.7%)
	Missing	1,285 (0.9%)
Number of Diseased Coronary Vessels	None	101 (0.1%)
	One	5,470 (3.9%)
	Two	27,177 (19.2%)
	Three	107,639 (76.2%)
	Missing	960 (0.7%)
Left Main Disease > 50%	No	47,223 (33.4%)
	Yes	45,275 (32.0%)
	Missing	48,849 (34.6%)
Ejection Fraction (%)	Median (IQR)	55.0 (45.0, 60.0)
	Missing	4,202 (3.0%)
Aortic Stenosis	No	134,339 (95.0%)
	Yes	4,241 (3.0%)
	Missing	2,767 (2.0%)
Mitral Stenosis	No	137,822 (97.5%)
	Yes	718 (0.5%)
	Missing	2,807 (2.0%)
Tricuspid Stenosis	No	137,969 (97.6%)
	Yes	90 (0.1%)
	Missing	3,288 (2.3%)
Pulmonic Stenosis	No	136,708 (96.7%)
	Yes	31 (0.0%)
	Missing	4,608 (3.3%)
Aortic Insufficiency	None	91,481 (64.7%)
	Trivial	13,594 (9.6%)
	Mild	10,818 (7.7%)
	Moderate	2,102 (1.5%)
	Severe	87 (0.1%)
	N/A or Not Documented	22,089 (15.6%)

	Effects	Overall N=141347
	Missing	1,176 (0.8%)
Mitral Insufficiency	None	44,494 (31.5%)
	Trivial	34,324 (24.3%)
	Mild	33,288 (23.6%)
	Moderate	8,895 (6.3%)
	Severe	658 (0.5%)
	N/A or Not Documented	18,734 (13.3%)
	Missing	954 (0.7%)
Tricuspid Insufficiency	None	46,845 (33.1%)
	Trivial	40,255 (28.5%)
	Mild	26,014 (18.4%)
	Moderate	4,185 (3.0%)
	Severe	356 (0.3%)
	N/A or Not Documented	22,349 (15.8%)
	Missing	1,343 (1.0%)
Pulmonic Insufficiency	None	74,106 (52.4%)
	Trivial	21,013 (14.9%)
	Mild	6,660 (4.7%)
	Moderate	568 (0.4%)
	Severe	46 (0.0%)
	N/A or Not Documented	37,276 (26.4%)
	Missing	1,678 (1.2%)

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

We used the same dataset of isolated CABG operations from October 2014 to September 2015 for the entire report. The three exceptions are:

1. For validity testing and the comparison of participants over time, we used STS participants with procedures during both October 2013 - September 2014 and October 2014 - September 2015 time periods.
2. For the analysis of population disparities, current and over time, we used eligible patients from STS participants with procedures between October 2011 and September 2015 and defined relevant subgroups by age, gender, race, ethnicity and insurance status.
3. For the analysis on the impact of exclusions, we included the cases with documented reasons for not having IMA used in CABG and/or with previous CABG.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We report trends of IMA use in CABG among the following groups: Age (<75, ≥75), Gender, Race (White, Black and Other), Hispanic Ethnicity and Insurance (<65, ≥65).

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

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

Reliability is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. For this NQF submission, the measurement of interest is each participant's observed proportion. The true value is the proportion that would be observed hypothetically if the sample size was very large (i.e. infinite).

For the j-th participant, let n_j denote the number of eligible patients, let $y_j$denote the number of patients receiving beta-blockers, and let $\bar{y}_j = y_j/n_j$ denote the proportion of patients receiving beta-blockers. In addition, let μ_j denote the underlying true value of $\bar{y}_j$. To estimate reliability, we assumed the following hierarchical model for the data. At the first stage of the hierarchy, we assume that y_j is distributed according to a binomial distribution with sample size n_j and probability parameter μ_j. At the second stage of the hierarchy, we assumed that μ_j varies across participants according to a Beta distribution with mean $E[\mu_j] = \alpha/(\alpha+\beta)$ and $\text{var}[\mu_j] = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$, where α and β are unknown parameters to be estimated from the data. The unknown parameters α and β were estimated via maximum likelihood using the BETABIN macro for SAS software (BETABIN, version 2.2, 2005. Qi Statistics). The sample for this analysis included all **1,041 participants** and **141,347 eligible patients** in the main study period October 2013-September 2014. After estimating α and β, we then calculated the reliability that would be achieved if the measure were to be calculated on a sample size of 30 patients per participant. This estimated reliability was calculated as

$$\text{reliability} = [\text{corr}(\bar{y}, \mu)]^2 = \frac{1}{1 + (\hat{\alpha} + \hat{\beta})/n}$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote maximum likelihood estimates of α and β, respectively, and n =30. Because reliability increases with n, and because the vast majority of STS participants have >30 eligible patients per year, the reliability calculated with n =30 patients per participant provides a conservative lower bound for the actual reliability that will be achieved when the measure is applied to STS data from a 1 year period. Using the above formula, we also calculated the sample size n required per participant to achieve reliability of at least 0.50, 0.60, and 0.70, and the proportion of STS participants with at least this number of eligible patients in the most recent 1-year testing sample.

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

Estimated parameter values of the beta distribution were $\hat{\alpha}$= 53.2907 and $\hat{\beta}$= 0.7503. The estimated reliability with 30 eligible patients per participant was 1/(1+ (53.2907 + 0.7503)/30)=0.36.

Based on these estimated parameter values, a sample size of 54 eligible patients per participant is needed to attain reliability of 0.50 and a sample size of 126 eligible patients per participant is needed to attain reliability of 0.70. During October 2014-September 2015, 80% of STS participants met the minimum required sample size for 0.50 reliability.

	Reliability	**Reliability**	**Reliability**

	0.50	0.60	0.70
Minimum required sample size per participant	54	81	126
Percent of participants meeting minimum sample size	80%	65%	41%

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

The proposed measure has adequate statistical reliability to be used for confidential feedback reporting as well as public reporting.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Critical data elements

Participating sites are randomly selected for participation in STS Adult Cardiac Surgery Database Audit, which is designed to evaluate the accuracy, consistency, and comprehensiveness of data collection and ultimately validate the integrity of the data contained in the database. Telligen has conducted audits on behalf of STS since 2006. In 2015, 10% of STS Adult Cardiac Surgery Database participants (N=107) were audited. The audit process involves re-abstraction of data for 20 cases and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each of the 82 variables, each variable category and overall. In 2015 the overall aggregate agreement rate was 96.17%, demonstrating that the data contained in the STS Adult Cardiac Surgery Database are both comprehensive and highly accurate.

Performance measure score

We calculated and compared the observed proportions of patients receiving the measure in the three performance groups. The measure has good face value if the three groups have different proportions as expected.

Face validity also implies that the measure is regarded as useful and valid by its intended users, including providers, consumers, payers, and regulators. The measure was developed with a panel of surgeon experts and statisticians. We have had near-universal acceptance of this composite by all stakeholders, with few if any relevant suggestions for change.

In addition, we tested the predictive validity of the measure. Predictive validity means that the results of this measure are predictive of future performance. We assessed the extent to which performance on this STS measure remains stable over time. In other words, does the measure at one point in time accurately predict performance at some later time?

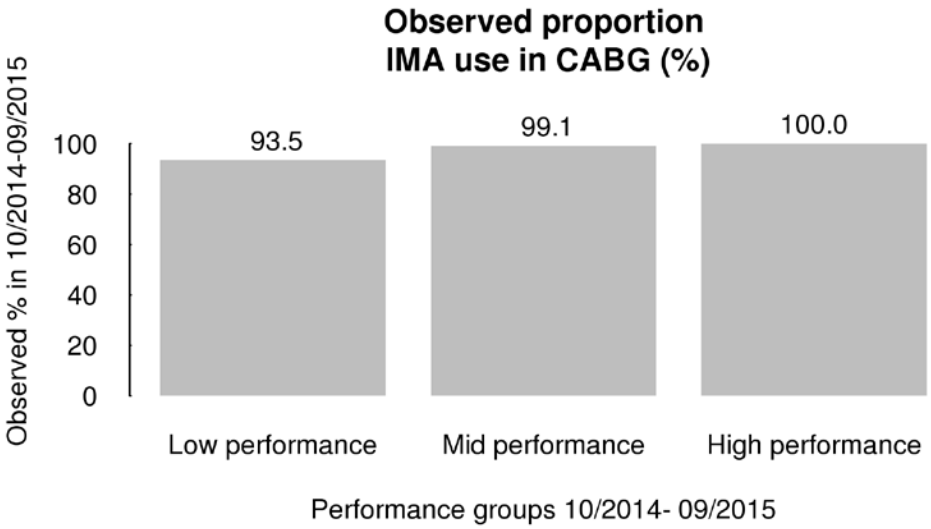
The tests on validity used the concept of performance outliers to be more formally introduced in 2b5: Participants were labeled as "low performance" if the 95% exact binomial confidence interval of its event rate lies entirely below the population average (in other words, the upper bound of the 95% CI < population

average). Participants were labeled as "high performance" if the 95% confidence interval lies entirely above 1. The remaining participants were labeled mid performance.

For each of the performance groups from the earlier period, we calculated the group specific measure proportions in the later period.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

STS participants deemed high performers by this measure have (on average) high rates of IMA use in CABG. Thus, differences in performance were clinically meaningful as well as statistically significant. This is illustrated in the figure below using data from October 2014 to September 2015. Compared to participants who were deemed as having lower than average performance, those with better-than-average performance had higher rate of IMA use in CABG (100.0% vs. 93.5%).

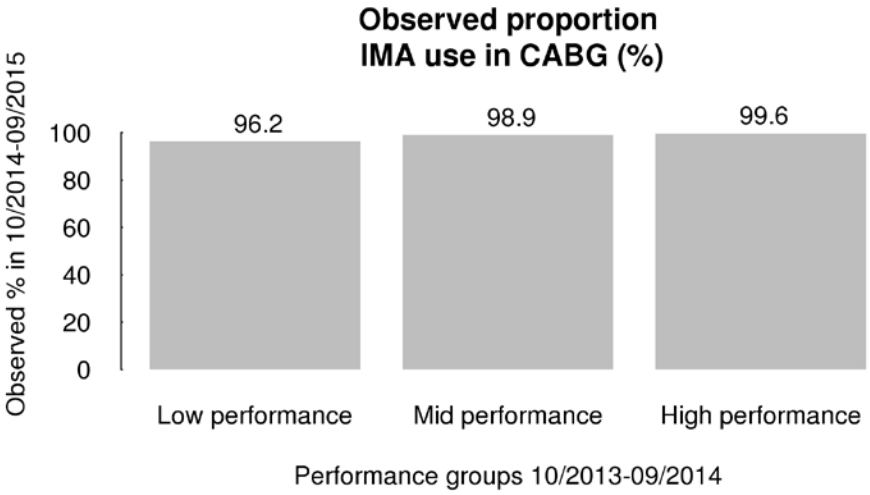


The predicted validity analysis was restricted to a sample of 1,013 STS participants with patients receiving the measure in both time periods: October 2013 - September 2014 and October 2014 - September 2015. Among participants who were high performance centers in October 2013-September 2014, 21.1% of them were also high performers for October 2014 - September 2015. For comparison, only 1.6% of participants who were mid performers in October 2013-September 2014 became high performers in October 2014 - September 2015. Thus, participants who performed better than average in October 2013-September 2014 were over 12 times more likely to be identified as better performers in the next year. Similarly, participants who were low performance entities in the early year were more likely to remain low performers in the later year. 2 participants jumped from low to high performing status (or vice versa) between the two adjacent 12-month periods. Thus, a consumer may reasonably expect that a high or low performer will likely be the same or became average in the near future, and a mid-performer is likely to remain average.

Change in performance categories between two time periods

		10/2014 – 09/ 2015		
		Low performance	Mid performance	High performance
10/2013 - 09/2014	Low performance	26	42	2
	Mid performance	46	863	15
	High performance	0	15	4

For each of the performance groups in the earlier period, we also calculated its aggregated proportion of patients receiving the measure in the later period. The aggregated proportions in the later periods were 99.6%, 98.9%, and 96.2% for the high, mid and low performance groups from the earlier period.



2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show that the measure reflects the proportion of patients with IMA use in CABG as designed, and that the past measure can be used to predict future performance. Together with face value, they support the validity of the measure.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded from the analysis cases with documented reasons for not having IMA used in CABG and/or with previous CABG. We believe these are clinically appropriate exclusions and are necessary to make the measure a consistent performance measure for the comparison across participants. The exclusions are precisely defined and specified.

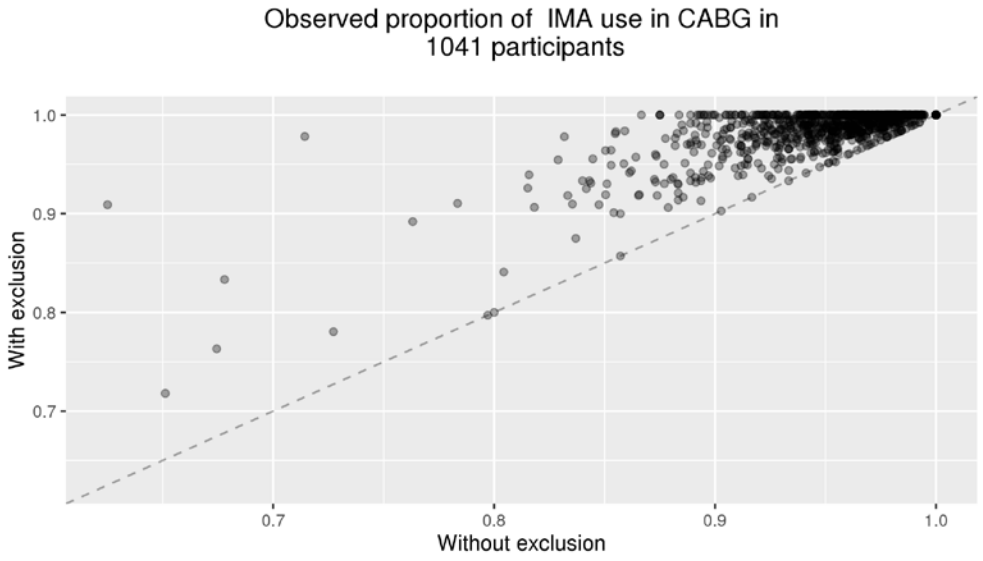
To show the impact of these exclusions, and how the measure would be distributed without it, we calculated and compared the distributions of the measure with and without the current exclusion criteria.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1041	1041
# Operations	141347	147965
Mean	0.99	0.95
STD	0.026	0.043
IQR	0.019	0.046
0%	0.72	0.62
10%	0.96	0.90
20%	0.98	0.93
30%	0.98	0.94
40%	0.99	0.95
50%	0.99	0.96
60%	1.00	0.97
70%	1.00	0.98
80%	1.00	0.98
90%	1.00	1.00
100%	1.00	1.00
Low performance	76, 7.3%	111, 10.7%
Mid performance	944, 90.7%	810, 77.8%
High performance	21, 2.0%	120, 11.5%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.60. The Pearson correlation is 0.73.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data*

*collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

For the measure to consistently quantify the quality per its definition, it is necessary to exclude cases if there were documented reasons for not having IMA used in CABG or with previous CABG. It has an impact on the results for many participants, and the results would be distorted without these appropriate exclusions.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** 42T **risk categories**
☐ **Other,** 42T

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)*

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The summary statistic provided is the participant's observed proportion of eligible patients with IMA use in CABG.

The degree of uncertainty surrounding an STS participant's IMA use in CABG measure estimate is indicated by the 95% exact binomial confidence interval (CI) of its observed proportion. Point estimates and CI's of the observed proportion for an individual STS participant are reported along with a comparison to the STS average proportion of the study time period. A performance category interpretation is also given to STS participants. **Since higher value indicates better performance**, an STS participant is designated as having higher/lower than average performance for the measure if the 95% CI lies entirely **above/below** the STS average. The remaining participants are labeled as not distinguishable from the STS average performance. For the simplicity of this report, we call the three groups 'high performance', 'low performance' and 'mid performance', respectively.

The method is equivalent to performing an exact binomial test with the null hypothesis that the participant has the same proportion of patients receiving the measure as the population average. Those with a test p-value smaller than 0.05 are the low and high performance groups.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

As shown in the table below, the proportion of STS ACSD participants performing better and worse than STS average has remained similar over the last two 12-month periods. On average, more than 90% of the participants have performance indistinguishable from the STS average, and the remaining participants have performed differently.

Distribution	10/2013 – 09/2014 Observed Proportion	10/2014 – 09/2015 Observed Proportion
# Participant	1056	1041
# Operations	140,354	141,347
Low performance	74, 7.0%	76, 7.3%
Mid performance	963, 91.2%	944, 90.7%
High performance	19, 1.8%	21, 2.0%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The statistical test and the construction of confidence interval are widely used and accepted. The participants identified as having performed differently from the average likely have true performance characteristics that are different. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the amount of outliers the measure detects.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Due to great data quality, the source fields required by the IMA use in CABG measure had only 0.1% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with 5% or more missing data are excluded from the measure calculation.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Overall, less than 0.1% of data were missing. 99% of participants had missing rate of 4% or lower. Seven out of 1048 participants were not included due to having missing rates higher than 5%.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.e., *what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The rates of missing data in the STS Adult Cardiac Surgery Database were very low and are getting lower. We therefore concluded that systematic missing data did not lead to bias in our measure.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

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

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

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those

whose performance is being measured.
The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Data Collection:
There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Planned	Current Use (for current use provide URL)
	Public Reporting This measure is one of eleven component measures of the STS CABG Composite Score. Approximately 49.8% of STS Adult Cardiac Surgery Database participants are voluntarily enrolled in the STS public reporting program. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online and Consumer Reports Health: www.ConsumerReports.org/hospitalratings Payment Program This is PQRS measure #43. The STS National Database was once again designated a Qualified Clinical Data Registry (QCDR) for PQRS reporting in 2016. STS reports this measure to CMS on behalf of all consenting surgeons. http://www.sts.org/quality-research-patient-safety/quality/physician-quality-reporting-system Quality Improvement with Benchmarking (external benchmarking to multiple organizations) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database

	Quality Improvement (Internal to the specific organization) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database

4a.1. **For each CURRENT use, checked above, provide:**

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

Please see table above

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Please see sections 1b.2 and 1b.4

In the table below we provide the overall trend over time of the measure performance at the patient level. The aggregate proportion of eligible patients with IMA use in CABG was computed for each time period. Although measure performance was above 98% during the entire period, we can see a moderate increase on the proportion of patients with IMA use in CABG over the last 4 years.

	10/2011 - 09/2012			10/2012 - 09/2013	10/2013 – 09/2014	10/2014-09/2015
All	98.36%	98.38%	98.66%	98.81%		

Geographic area and number and percentage of accountable entities and patients included
Number of participants and operations by geographic regions, during the two last consecutive time periods, October 2013-September 2014 and October 2014-September 2015.

	10/2013 – 09/2014				10/2014 – 09/2015			
	Midwest	Northeast	South	West	Midwest	Northeast	South	West
# Participant	305	133	397	221	297	136	393	215
% Participant	28.9%	12.6%	37.6%	20.9%	28.5%	13.1%	37.8%	20.7%
# Operation	34493	22337	61104	22420	34528	22655	61382	22782
% Operation	24.6%	15.9%	43.5%	16.0%	24.4%	16.0%	43.4%	16.1%

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of**

high-quality, efficient healthcare for individuals or populations.
N/A

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2015, 10% of participants, i.e., 107 facilities, were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

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

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0116 : Anti-Platelet Medication at Discharge
0117 : Beta Blockade at Discharge
0118 : Anti-Lipid Treatment Discharge
0119 : Risk-Adjusted Operative Mortality for CABG
0127 : Preoperative Beta Blockade
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** 0134_Use_of_IMA_in_CABG_Appendix_-_S.9-_1b.2-_1b.4-_Guidelines.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
- Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Chair, Workforce on National Databases; surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2004
Ad.3 **Month and Year of most recent revision:** 06, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0706
Measure Title: Risk Adjusted Colon Surgery Outcome Measure
Measure Steward: American College of Surgeons
Brief Description of Measure: This is a hospital based, risk adjusted, case mix adjusted morbidity and mortality aggregate outcome measure of adults 18+ years undergoing colon surgery.
Developer Rationale: Reduced morbidity and mortality rates following colon surgery.

Numerator Statement: The outcome of interest is 30-day, hospital-specific risk-adjusted (all cause) mortality, unplanned reoperation, 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). All outcomes are definitively resolved within 30 days of any ACS NSQIP listed (CPT) surgical procedure. All variables (fields) are explicitly defined in the tradition of the ACS NSQIP and definitions are also submitted in these materials. The original endorsed measure included venous thromboembolism (VTE) as eligible morbidity events, including deep venous thrombosis requiring therapy and pulmonary embolism.

The current set of mortality and major complications for this measure was chosen based on prior work revealing that these complications are related to other important criteria such as large contributions to excess length of stay, large complication burdens, or correlations with mortality. (Merkow et al. 2013) In addition, the desire to limit the outcomes to significant events (ie- some degree of severity according to certain criteria) is the reason that superficial wound infection is excluded from the measure. The current submission removes VTE from the measure as recent publications have demonstrated it is highly subject to surveillance bias. A recent study of 2,838 hospitals found that increased VTE prophylaxis adherence was associated with worse risk-adjusted VTE event rates. (Bilimoria 2013 JAMA) Paradoxically hospitals with higher quality, identified by number of accreditations and quality initiatives, had worse VTE rates. The explanation for this paradoxical relationship is suggested by the association of higher rates of VTE imaging studies among these hospitals with higher rates of VTE detection. (Bilimoria, Chung et al. 2013, Ju, Chung et al. 2014, Chung, Ju et al. 2015)

Bilimoria, K. Y., J. Chung, M. H. Ju, E. R. Haut, D. J. Bentrem, C. Y. Ko and D. W. Baker (2013). "Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure." Jama 310(14): 1482-1489.
Chung, J. W., M. H. Ju, C. V. Kinnier, M. W. Sohn and K. Y. Bilimoria (2015). "Postoperative venous thromboembolism outcomes measure: analytic exploration of potential misclassification of hospital quality due to surveillance bias." Ann Surg 261(3): 443-444.
Ju, M. H., J. W. Chung, C. V. Kinnier, D. J. Bentrem, D. M. Mahvi, C. Y. Ko and K. Y. Bilimoria (2014). "Association between hospital imaging use and venous thromboembolism events rates based on clinical data." Ann Surg 260(3): 558-564; discussion 564-556.
Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the american college of surgeons colectomy composite outcome quality measure. Ann Surg. 2013;257(3):483-489.

Denominator Statement: Patients undergoing any ACS NSQIP listed (primary CPT) colon procedure. (44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44160, 44204, 44205, 44206, 44207, 44208, 44210)
Denominator Exclusions: As noted above, cases are collected so as to match ACS NSQIP inclusion and exclusion criteria, thereby permitting valid application of ACS NSQIP model-based risk adjustment. Therefore, trauma and transplant surgeries are excluded as are surgeries not on the ACS NSQIP CPT list as eligible for selection (see details in next item). Patients who are ASA 6 (brain-death organ donor) are not eligible surgical cases. Of note, the measure excludes patients identified as having had prior surgical procedures within 30 days of a potential index procedure, since this measure is based on 30 day outcomes. A patient who is identified as having had a prior surgical procedure within 30 days of the index case being considered is excluded from accrual. A patient who has a second surgical procedure performed within 30 days after an index procedure has the second procedure recorded as a "Return to the operating room within 30 days" (one of the outcomes defined), but the second procedure cannot be accrued into the program as a new index procedure.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Registry, Management Data, Paper Medical Records
Level of Analysis: Facility, Population : National

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

Maintenance of Endorsement – Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This measure was last reviewed in 2012, is for hospital based, risk adjusted, case mix adjusted morbidity and mortality aggregate outcome measure of adults 18+ years undergoing colon surgery.
- The rationale provided to support this measure states that inpatient colorectal procedures, compared to other general surgeries, are responsible for a significant proportion of adverse events and excessive length of stay. Data from the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) colorectal risk calculator (based on 2006-2007 NSQIP data) showed overall morbidity for colorectal surgery at 24.3%, serious morbidity at 11.4% and mortality at 3.9%.
- The developer states that hospital level improvements in quality of care can influence postoperative outcomes in colorectal surgery regardless of hospital volume and that an outcome measure that includes serious morbidity and mortality for patients undergoing colorectal surgery is valid and feasible.

Question for the Committee:

- *Does the Committee agree the underlying rationale for the measure remains reasonable?*
- *Does the Committee agree that the evidence supports the measure?*
- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm: Health outcome (Box 1)→relationship between outcome and at least one

healthcare action identified/supported by stated rationale (Box 2) →Pass

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

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

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

- The ACS National Surgical Quality Improvement Program (NSQIP) collects detailed clinical data and provides reports to participating hospitals with risk-adjusted comparisons with a surgical quality standard.
  - A recent analysis of NSQIP showed that over 8 years in the program, 62% hospitals improved their performance in mortality and 71% of hospitals improved their performance in risk-adjusted complications.
  - Further, the developer reports that hospitals may observe annual reductions in mortality and morbidity, at 0.8% and 3.1% , respectively.
  - The developer reports that for 2014, the observed to expected (O/E) ratios for colon surgery mortality and serious morbidity range from 0.66 to 1.4 for participating hospitals and the Interquartile range was 0.15

Disparities

The developer also reports disparities in race, age, and socioeconomic factors identified for colorectal surgery:

- Black patients (41.3%), compared with white patients (45.4%), have lower 5-year overall survival rates after surgery for colon cancer and are less likely to receive adjuvant therapy after rectal cancer resection (48.6% versus 60.9%)
- Compared to non-Hispanic whites, blacks, American Indians, Chinese, Filipinos, Koreans, Hawaiians, Mexicans, South/Central Americans, and Puerto Ricans are 10-60% more likely to be diagnosed with late stage colorectal cancer.
- More recently, Hardiman et al. demonstrated that of patients with primary colon tumors, those at least 80 years were less likely to have colectomy, have fewer lymph nodes removed, and less likely to receive chemotherapy for every stage than those who were younger than 80 years.
- Frederiksen et al. found that postoperative mortality after elective colorectal cancer surgery was significantly lower in patients with higher incomes and education and among those who owned homes.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1a.

- good composite outcome measures
- Why not keep PE in, even if eliminating DVT?
- Does UTI exert an undue influence on the outcomes? Would a center with low death rate but high UTI be better or worse than one with high death rate but low UTI?
- Only 10 high and 12 low outliers among 450 hospitals after risk adjustment
- Are calculations based on NSQIP sample, or all eligible cases? How similar are these populations?

1b.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic Clinical Data, Electronic Health Record, Imaging/Diagnostic Study, Laboratory, Electronic , Clinical Data : Registry, Management Data, Paper Medical Records

Specifications:

- The measure is specified as a facility and population (national)-level measure for the ambulatory care (ambulatory surgery center) and hospital/acute care setting.
- The numerator includes 30-day, hospital-specific risk-adjusted (all cause) mortality, unplanned reoperation, or a number of morbidities. All outcomes are definitively resolved within 30 days of any ACS NSQIP listed (CPT) surgical procedure.
- The denominator includes patients undergoing any ACS NSQIP listed (primary CPT) colon procedure.
- Trauma and transplant surgeries are excluded as well as surgeries not on the ACS NSQIP CPT list as eligible for selection. Additional denominator exclusions are detailed here.
- This outcome measure is risk adjusted, using a statistical risk model. It is unclear which risk factors are included in this measure.

Questions for the Committee :

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

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Previous reliability testing included an examination of the variation between hospitals from the ACS NSQIP database for 2008 using the intra-class correlation coefficient (ICC) .

Describe any updates to testing:

- Score level testing using a method described as a standard method and equations derived from the 2011-2014 data set (2014, 451 hospitals, 33,747 cases).

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Reliability was assessed using information from a random intercept, fixed slope, hierarchical logistic regression model and applying the Spearman Brown prophecy formula.

Results of reliability testing

- Reliability was evaluated for 451 hospitals collecting data during 2014. The mean number of cases per hospital in the 2014 data set was 74.8, but there was a positive skew in the distribution of sample sizes (median =51).
- The developer 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 recommended for this measure).

- Based on a regression analysis, the developer determined that an empirical estimate of the sample size required to achieve reliability of 0.4 is 99. The developer notes that this number is considered "an appropriate and achievable target for the number of colon cases for hospitals interested in participating in this measure given current case numbers. Nevertheless, hospitals with fewer cases might still benefit from participation in this measure." The reliability of the 2014 dataset was examined under 4 conditions: VTE (included or not included) and SES variables (included or not included). Over 30% of hospitals had a "minimally acceptable" reliability estimate for the measure across all 4 variations of VTE and SES variables (see below):

Calibration range	Percent without VTE+, SES-
0.00-0.20	35.03
0.21-0.40	33.48
0.41-0.60	24.39
0.61-0.80	7.10

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the method of testing and interpretation of results clear and compelling?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empiric reliability testing (Box 2) → Testing at measure score (Box 4) → Method described (Box 5) → Confidence in measure score →moderate

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**
Specification not completely consistent with evidence

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- In the previous submission, the measure was tested using C-statistics and the Hosmer-Lemeshow P value statistic.

Describe any updates to testing:

- The developer describes the risk model diagnostics approach to evaluation of risk model quality and processes. While important information, the demonstration of risk model adequacy does not meet NQF requirements for validity testing of the measure score.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- N/A

Validity testing results:

- N/A

Questions for the Committee:

o *What approach would the Committee accept to receive and evaluate validity testing in the current cycle?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer notes patients are excluded from the measure for the following reasons:
  - Surgeries not on the ACS NSQIP CPT list
  - Trauma and transplant surgeries
  - ASA 6 (brain-death organ donor)
  - Prior surgical procedures within 30 days of a potential index procedure

  Information regarding number of excluded cases is not provided, thus analysis of exclusions cannot be appreciated.

Questions for the Committee:

o *Are the exclusions consistent with the evidence?*

o *Are any patients or patient groups inappropriately excluded from the measure?*

o *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

- Statistical risk modeling is performed in a step-wise fashion – case mix adjustment, variable selection, then risk adjustment.
- For variable selection of risk factors, logistic regression performed using NSQIP predictors demonstrating statistical significance (P<.05) from which a subset is chosen to create a predictor set. The original set of predictors is included. For this measure the 6 predictors are: ASA class, CPT risk, functional status, operative indication, emergency case and wound class.
- Additional variables modeled to explore SDS were race, Hispanic ethnicity, and race. Inclusion and exclusion of VTE was also assessed.
  - The developer noted that the addition of SES factors was not influential in risk adjustment (weighted kappa=1) and that removal of VTE has an important effect on outlier and decile status (weighted kappa=0.765).
- The developer reports that C-statistic has been used to evaluate discrimination of the risk model and is applied to evaluation of exclusion/inclusion of VTE and SDS factors in the current measure. Other tests to demonstrate accuracy of prediction about probability (Brier), and goodness of fit for models assessing observed and expected rates (Hosmer-Lemeshow), were computed for the 2011-2014 dataset
  - The developer reports that model quality remains consistent when the 2011-2014 equations are applied to the 2010 dataset.

Questions for the Committee:

- *Is the risk-adjustment strategy included in the measure clear, complete and appropriate for the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Is there a conceptual rationale to include (or not include) SDS factors in the risk adjustment approach? If so, are the relevant risk factors considered?*
- *Do you agree with the developer's decision not to include SDS factors in the risk-adjustment approach?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The default methodology for discrimination performance is based on the computed 95% CI (using Ulm's method) for the O/E ratio. If the interval is entirely above1.0, the hospital is identified as having performance significantly worse than expected. If the interval is entirely below 1.0, the hospital is identified as having performance significantly better than expected. If the interval overlaps 1.0 the hospital is performing "as expected."
- Using a 95% confidence interval for the observed to expected events ration, the original surgery outcome measure (without SES and with VTE) identified 12 low and 10 high outliers among the 451 hospitals with 2014 data.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Participating hospitals collect and report data to ACS NSQIP. Hospitals that do not participate can submit data to the implementing organization.

2b7. Missing Data

The developer reports high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Guidance from the Validity Algorithm:

Specifications consistent with the evidence (Box 1) → Potential threats to validity addressed (Box 2)

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

Preliminary rating is based on lack of clarity about risk adjustment strategy, testing, and comparability of data sources/methods

Committee pre-evaluation comments

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

Criterion 3. Feasibility

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

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

The developer noted the following:

- Data is generated as byproduct of care processes during care delivery and coding/abstraction is performed by someone other than person obtaining original information; coding/abstraction is performed by dedicated abstraction personnel.
- No data elements are in defined fields in electronic sources.
- A completely electronic medical record (EMR) would be needed to capture all risk factors that enter into the model. In addition, a software module (currently available to ACS NSQIP subscribers) will be required to transfer

information from the EMR to a measure submission database. The ACS NSQIP is in the process of developing an automated process with EMR vendors, however, electronic entry for this measure is not currently available.
- ACS NSQIP has been open to subscription by private sector hospitals since 2004 and over 600 hospitals participate.

Questions for the Committee:

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

o *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

o *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

Public Reporting (Hospital Compare) – 131 hospitals currently report their risk-adjusted surgery outcomes data for NQF endorsed measures from ACS.

Quality Improvement (both internal and external with benchmarking) – The developer reports there are over 600 hospitals currently participating in ACS NSQIP and receiving risk-adjusted benchmarking reports. ACS NSQIP hospitals utilize their internal data for the purpose of quality improvement initiatives specific to the organization.

Improvement results

- Previously described in 2b5 above.

Unexpected findings (positive or negative) during implementation

- The developer does not report any unexpected findings

Potential harms

- The developer does not report any potential harms

Feedback : N/A

Questions for the Committee:

o *How can the performance results be used to further the goal of high-quality, efficient healthcare?*

o *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ High ☒ Moderate ☐ Low ☐ Insufficient
Committee pre-evaluation comments **Criteria 4: Usability and Use**

Criterion 5: Related and Competing Measures
Related or competing measures 0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB). 0697 : Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure **Harmonization** The developer reports that the measure specifications are completely harmonized with the related measures.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

NQF #: 0706 NQF Project: Patient Outcomes Measures: Phases I and II

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT
Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence. *Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.* (evaluation criteria)
1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*: This is a risk-adjusted outcome measure. Approximately 500,000 inpatient colorectal procedures are performed annually in the United States and, compared to other general surgery procedures, colon surgery is responsible for a significant proportion of adverse events and excessive length of stay.(Schilling, Dimick et al. 2008) Much of the excess length of stay (LOS), charges, and death can be attributed to postoperative complications.(Kirkland et al. 1999, Short, Aloia et al. 2014, Knechtle, Perez et al. 2015) The LOS for patients who have any postoperative complication is 3–11 days longer than that for patients who do not experience complications.(Kirkland et al. 1999, Coello et al. 2005) In an analysis of data from the Veterans Administration (VA) NSQIP, the occurrence of a complication 30 days in duration reduced median patient survival by 69% independent of preoperative patient risk. (Khuri et al. 2005) Higher hospital volume has not been correlated with improved outcomes for colorectal surgery as is the case for pancreatic and esophageal surgeries.(Birkmeyer et al. 2006, Finlayson et al. 2003) Nevertheless, higher surgeon volume has been correlated with improved outcomes with lower volume surgeons benefiting from performance of colorectal procedures at high volume hospitals.(Karanicolas et al. 2008, Harmon et al. 1999) These findings suggest that hospital level improvements in quality of care can influence postoperative outcomes in colorectal surgery regardless of hospital volume.

Ultimately, a variety of process measures exist that correlate to improved outcomes in colorectal surgery but there are no guidelines tied directly to outcome measurement. The NSQIP program has shown that a composite outcome that includes serious morbidity and mortality for patients undergoing colorectal surgery is valid and feasible to measure. (Merkow et al. 2013) Current reports provided back to individual hospitals based on colorectal surgical models include 30-day mortality, 30-day morbidity, length of stay, and surgical site infections.

Birkmeyer, J.D., et al., Volume and process of care in high-risk cancer surgery. Cancer, 2006. 106(11): p. 2476-81.

Coello, R., et al., Adverse impact of surgical site infections in English hospitals. J Hosp Infect, 2005. 60(2): p. 93-103.

Finlayson, E.V. and J.D. Birkmeyer, Effects of hospital volume on life expectancy after selected cancer operations in older adults: a decision analysis. J Am Coll Surg, 2003. 196(3): p. 410-7.

Harmon, J.W., et al., Hospital volume can serve as a surrogate for surgeon volume for achieving excellent outcomes in colorectal resection. Ann Surg, 1999. 230(3): p. 404-11; discussion 411-3.

Karanicolas, P.J., et al., The more the better? The impact of surgeon and hospital volume on in-hospital mortality following colorectal resection. Ann Surg, 2009. 249(6): p. 954-9.

Khuri, S.F., et al., Determinants of long-term survival after major surgery and the adverse effect of postoperative complications. Ann Surg, 2005. 242(3): p. 326-41; discussion 341-3.

Kirkland, K.B., et al., The impact of surgical-site infections in the 1990s: attributable mortality, excess length of hospitalization, and extra costs. Infect Control Hosp Epidemiol, 1999. 20(11): p. 725-30.

Knechtle, W. S., S. D. Perez, R. L. Medbery, B. D. Gartland, P. S. Sullivan, S. J. Knechtle, D. A. Kooby, S. K. Maithel, J. M. Sarmiento, V. O. Shaffer, J. K. Srinivasan, C. A. Staley and J. F. Sweeney (2015). "The Association Between Hospital Finances and Complications After Complex Abdominal Surgery: Deficiencies in the Current Health Care Reimbursement System and Implications for the Future." Ann Surg 262(2): 273-279.

Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the american college of surgeons colectomy composite outcome quality measure. Ann Surg. 2013;257(3):483-489.

Schilling, P. L., J. B. Dimick and J. D. Birkmeyer (2008). "Prioritizing quality improvement in general surgery." J Am Coll Surg 207(5): 698-704.

Short, M. N., T. A. Aloia and V. Ho (2014). "The influence of complications on the costs of complex cancer surgery." Cancer 120(7): 1035-1041.

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

Observational evidence is based on prospectively collected rigorously controlled data (ACS NSQIP). As noted, a variety of processes influence outcomes but no guidelines address outcome measurement as a means toward quality improvement. McGlynn et al. showed stark deficits in the adherence to standard processes and quality of health care being delivered to the American public with only 55% of colorectal cancer patients receiving the recommended care.(McGlynn et al. 2003) The morbidity and mortality associated with colorectal operations are common and pose considerable risk to patients.(Schilling et al. 2008) Several statistical models have been proposed to predict surgical risk for these operations.(Bromage et al. 2007, Tekkis et al. 2004, Duval et al. 2006, Alves et al. 2007, Fazio et al. 2004) The development of the ACS NSQIP colorectal risk calculator stemmed from 2006-2007 NSQIP data demonstrating an overall morbidity rate for colorectal surgery at 24.3%, serious morbidity rate at 11.4%, and mortality rate at 3.9%. (Cohen et al. 2009)

Several Surgical Care Improvement Project (SCIP) measures were developed and tied to outcomes specific to colorectal surgery. SCIP measures focus on decreasing infections, specifically postoperative wound infections, urinary tract infections, and pneumonia. Despite identification of several process measures, adherence to these measures alone is rapidly achievable while complications and poor outcomes remain a significant issue. (Forbes et al. 2008, Hedrick et al. 2007) Ingraham et al. evaluated the relationship between SCIP process measures and ACS NSQIP outcomes and found that 15 out of 16 measures demonstrated non-significant associations with outcomes. The exception was that administration of appropriate antibiotics demonstrated a significant relationship with reduction in SSI.(Ingraham, Cohen et al. 2010) Appropriate antibiotic use in colorectal surgery has been studied in several randomized controlled trails and meta-analysis.(Woodfield et al. 2009) The lack of correlation between SCIP process measures and meaningful clinical

outcomes underscores the importance of tracking clinical outcomes as an end point in quality improvement. Measures such as the colon surgery death or serious morbidity outcome provide a clinically meaningful endpoint for patients and providers to follow.

There are few randomized controlled trials comparing overall outcomes in colorectal surgery, hence, there are few process measures developed specific to colectomy alone. This is evidenced further by two articles that provide a comprehensive review of colorectal surgical quality. (McGory et al. 2006, Gagliardi et al. 2005) In both articles, due to the lack of level 1/trial based evidence, expert opinion and a systematic review of the literature were used to identify candidate quality indicators. More specifically, Gagliardi et al. used a three-step modified Delphi approach, whereas McGory et al. used the RAND/University of California-Los Angeles appropriateness method. (Fink et al. 1984, Brook 1994) Nevertheless, these reviews of the best available literature focus on various process measures and do not provide feedback about a composite outcome measure based on colorectal surgery.

Gagliardi and colleagues identified 45 candidate indicators, of which 37 (82%) were considered valid by the panel. McGory and colleagues identified 142 candidate indicators, of which 92 (65%) were considered valid. In the former study, panelists were asked to rank candidate indicators; the investigators reported only the top 15 prioritized quality indicators as their final recommendation for improving the quality of colorectal cancer surgery. Although such reporting does present a parsimonious set of quality indicators, the detailed list of 92 quality indicators presented in the latter study is comprehensive and encompasses the entire perioperative time period, including preparation of the patient for surgery, intraoperative issues, and postoperative processes of care. In that regard, the focus of the quality indicators is an interesting difference between the two studies.

Gagliardi and colleagues focused on four outcome measures (e.g., 30-day mortality) and four measures evaluated at the province level (e.g., 5-year survival). In contrast, McGory and colleagues focused on process and structural measures and did not include any outcome measures. The latter investigators presented six quality domains spanning the 92 indicators: surgeon privileging (e.g., credentialing for laparoscopic colectomy), preoperative evaluation (e.g., staging), patient-provider discussions (e.g., informed consent), medications (e.g., antibiotic prophylaxis), intraoperative care (e.g., prevention of ureteral injury), and postoperative management (e.g., control of blood glucose). In addition to focusing on process rather than outcomes, all of their measures were recorded at the provider level, not the hospital, county, or state level. It seems that the two studies developed indicators with different agendas in mind, and as such, the potential application of these studies for quality improvement will be broad.

Though comprehensive, these reviews of available literature focus on processes of care in order to influence outcome. A Delphi-based consensus process was conducted by the American Society of Colon and Rectal Surgeons, the American Board of Colon and Rectal Surgeons and the Colo-Rectal Education System Template Committee in order to develop key endpoints for quality improvement.(Manwaring, Ko et al. 2012) There were 89 process and outcome measures identified, of which tracking complications and mortality remained highly scored across all categories.

Risk-adjusted, benchmarked outcomes measures provide enormous motivation for hospitals to see their outcomes improve. A recent analysis indicates that over 8 years in the program, 62% and 71% of hospitals improve their performance in mortality and risk-adjusted complications.(Cohen, Liu et al. 2016) This is consistent with hospital improvements found in prior publications.(Hall, Hamilton et al. 2009) Hospitals may observe annual reductions of 0.8% in mortality and 3.1% in morbidity; though small, these reductions provide cumulative benefit as hospitals continue participation in the ACS NSQIP program. Other research has shown NSQIP improvements as well and many are referenced in the above citations.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles*):

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: RAND/UCLA appropriateness methodology. (Fink et al. 1984, Brook 1994)

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: No Level 1A/Randomized Controlled Trials for overall colorectal surgery outcomes. This measure is developed based on prospectively collected rigorously controlled data. Current literature is primarily level 2 and 3 evidence being validated by multiple expert panels with trials exploring several process related components.

1c.14 Summary of Controversy/Contradictory Evidence: Colon and rectal surgeries are a common procedure group that are performed at a high number of hospitals across the U.S., and are associated with significant morbidity and mortality. Current process measures do not correlate with clinical risk adjusted outcomes as discussed elsewhere in this measure proposal. Specific process measures for colectomy do not have high enough evidence base for accountability. With this measure, colorectal surgery quality of care is evaluated using a clinical, risk-adjusted outcomes measure. As described, several process measures exist and additional groups and panels have described outcomes of interest in colorectal surgery. While further studies are needed to identify additional indicators, no true controversy or contradictory evidence currently exists when observing a composite outcome that spans a wide array of clinically significant outcomes. The provision of risk-adjusted, benchmarked outcomes data provides significant impetus for quality improvement to participating ACS NSQIP hospitals.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Alves, A., et al., The AFC score: validation of a 4-item predicting score of postoperative mortality after colorectal resection for cancer or diverticulitis: results of a prospective multicenter study in 1049 patients. Ann Surg, 2007. 246(1): p. 91-6.

Brook, R.H., The RAND/UCLA Appropriateness Method, in Methodology perspectives, K.A. McCormick, S.R. Moore, and R.A. Siegel, Editors. 1994, Public Health Services, U.S. Department of Health and Human Services: Rockville, MD. p. 59-70.

Bromage, S.J. and W.J. Cunliffe, Validation of the CR-POSSUM risk-adjusted scoring system for major colorectal cancer surgery in a single center. Dis Colon Rectum, 2007. 50(2): p. 192-6

Cohen, M.E., et al., Development of an American College of Surgeons National Surgery Quality Improvement Program: morbidity and mortality risk calculator for colorectal surgery. J Am Coll Surg, 2009. 208(6): p. 1009-16.

Cohen et al. 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

Duval, H., et al., [The Association Francaise de Chirurgie (AFC) colorectal index: a reliable preoperative prognostic index in colorectal surgery]. Ann Chir, 2006. 131(1): p. 34-8.

Fazio, V.W., et al., Assessment of operative risk in colorectal cancer surgery: the Cleveland Clinic Foundation colorectal cancer model. Dis Colon Rectum, 2004. 47(12): p. 2015-24.

Fink, A., et al., Consensus methods: characteristics and guidelines for use. Am J Public Health, 1984. 74(9): p. 979-83.Forbes, S.S., et al., Implementation of evidence-based practices for surgical site infection prophylaxis: results of a pre- and postintervention study. J Am Coll Surg, 2008. 207(3): p. 336-41.

Gagliardi, A.R., et al., Development of quality indicators for colorectal cancer surgery, using a 3-step modified Delphi approach. Can J Surg, 2005. 48(6): p. 441-52.

Hall, B. L., B. H. Hamilton, K. Richards, K. Y. Bilimoria, M. E. Cohen and C. Y. Ko (2009). "Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals." Ann Surg 250(3): 363-376.

Hedrick, T.L., et al., Efficacy of protocol implementation on incidence of wound infection in colorectal operations. J Am Coll Surg, 2007. 205(3): p. 432-8.

Ingraham, A. M., M. E. Cohen, K. Y. Bilimoria, J. B. Dimick, K. E. Richards, M. V. Raval, L. A. Fleisher, B. L. Hall and C. Y. Ko (2010). "Association of surgical care improvement project infection-related process measure compliance with risk-adjusted outcomes: implications for quality measurement." J Am Coll Surg 211(6): 705-714.

Manwaring, M. L., C. Y. Ko, J. W. Fleshman, Jr., D. E. Beck, D. J. Schoetz, Jr., A. J. Senagore, R. Ricciardi, L. K. Temple, A. M. Morris and C. P. Delaney (2012). "Identification of consensus-based quality end points for colorectal surgery." Dis Colon Rectum 55(3): 294-301.

McGory, M.L., P.G. Shekelle, and C.Y. Ko, Development of quality indicators for patients undergoing colorectal cancer surgery. J Natl Cancer Inst, 2006. 98(22): p. 1623-33.

McGlynn, E.A., et al., The quality of health care delivered to adults in the United States. N Engl J Med, 2003. 348(26): p. 2635-45.
Schilling, P.L., J.B. Dimick, and J.D. Birkmeyer, Prioritizing quality improvement in general surgery. J Am Coll Surg, 2008. 207(5): p. 698-704.
Tekkis, P.P., et al., Development of a dedicated risk-adjustment scoring system for colorectal surgery (colorectal POSSUM). Br J Surg, 2004. 91(9): p. 1174-82.
Woodfield, J.C., N. Beshay, and A.M. van Rij, A Meta-Analysis of Randomized, Controlled Trials Assessing the Prophylactic Use of Ceftriaxone. A Study of Wound, Chest, and Urinary Infections. World J Surg, 2009.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
Similar to the quality indicators for patients undergoing colorectal surgical procedures (as described above), there are process-based guideline recommendations specifically for cancer operations for colorectal cancer, however, these recommendations do not address colectomy broadly nor are they outcomes based. There are no national clinical practice guidelines specific for colorectal surgical procedures that are applicable to the current maintenance of endorsement submission for this measure. (http://www.guideline.gov/; accessed 5/16/2016).

1c.17 Clinical Practice Guideline Citation: N/A

1c.18 National Guideline Clearinghouse or other URL: N/A

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: N/A

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: N/A

1c.24 Rationale for Using this Guideline Over Others: N/A

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: 1c.26 Quality: 1c.27 Consistency:

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0706_Evidence_Maintenance-May2016.doc

1b. Performance Gap

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

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

- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Reduced morbidity and mortality rates following colon surgery.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Morbidity and mortality rates vary by institution for colorectal surgery. Prior analysis indicated significant variability in overall morbidity after colorectal surgery, prompting the development of the colorectal surgery risk calculator.(Cohen et al. 2009) ACS NSQIP has been using similar O/E ratios to measure outcomes in the program since its inception in the VA in the 1990s. The success of this program and the satisfaction of participants provide evidence of success in achieving results with information similar to this outcome measure.

This risk-adjusted and benchmarked measure provides enormous motivation for hospitals to see their outcomes improve. A recent analysis indicates that over 8 years in the program, 62% and 71% of hospitals improve their performance in mortality and risk-adjusted complications. (Cohen, Liu et al. 2015) This is consistent with hospital improvements found in prior publications. (Hall, Hamilton et al. 2009) The effect on avoided complications is significant, as the analysis demonstrates that hospitals may observe annual reductions of 0.8% in mortality and 3.1% in morbidity; though small, these reductions provide cumulative benefit as hospitals continue participation in the ACS NSQIP program. Other research has shown NSQIP improvements as well- many are referenced in the above citations.

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. 2015) For 2014, there were 451 hospitals contributing 33,747 colorectal surgery cases. The O/E ratios for colon surgery mortality and serious morbidity range from 0.66 to 1.4 for participating hospitals, demonstrating a wide gap between those performing better and worse than expected after risk and case mix adjustment. The interquartile range for the O/E ratio is 0.15, and the 10th percentile and 90th percentile O/E ratios were 0.86 and 1.17, respectively.

Cohen, M.E., et al., Development of an American College of Surgeons National Surgery Quality Improvement Program: morbidity and mortality risk calculator for colorectal surgery. J Am Coll Surg, 2009. 208(6): p. 1009-16.
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
Hall, B. L., B. H. Hamilton, K. Richards, K. Y. Bilimoria, M. E. Cohen and C. Y. Ko (2009). "Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals." Ann Surg 250(3): 363-376.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Cohen, M.E., et al., Development of an American College of Surgeons National Surgery Quality Improvement Program: morbidity and mortality risk calculator for colorectal surgery. J Am Coll Surg, 2009. 208(6): p. 1009-16.
Cohen et al. 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

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
There is wide variation in the care and outcomes of patients undergoing colorectal resections in the U.S. In 2003 McGlynn et al. showed stark deficits in the adherence to standard processes and quality of health care being delivered to the American public with only 55% of colorectal cancer patients receiving the recommended care.(McGlynn et al. 2003) Obese patients have been shown to have higher rates of death due to cancer of the colon and rectum.(Calle et al. 2003) In patients undergoing rectal resections, obesity is associated with a higher anastomotic leakage rates (16% versus 6% for non-obese patients, P <0. 05).(Benoist et al. 2000)

Race also plays an important role in outcomes after colorectal resections. Black patients, compared with white patients, have lower

5-year overall survival rates after surgery for colon cancer (41.3% v 45.4%, respectively; P < .001) (Breslin et al. 2009) and are less likely to receive adjuvant therapy after rectal cancer resection (48.6% versus 60.9%, p < 0.0001). (Morris et al. 2006) Compared to non-Hispanic whites, blacks, American Indians, Chinese, Filipinos, Koreans, Hawaiians, Mexicans, South/Central Americans, and Puerto Ricans are 10-60% more likely to be diagnosed with late stage colorectal cancer. (Chien et al. 2005) Hardiman et al. demonstrated through a retrospective review of prospectively collected data on 10,433 patients diagnosed with primary colon tumors that individuals who were at least 80 years old were less likely to have colectomy for advanced or metastatic disease, have fewer lymph nodes removed, receive chemotherapy for every stage than those who were younger than 80 years old. (Hardiman et al. 2009) Disparities in socioeconomic factors have also been identified for colorectal surgery. A study of 7,160 patients from Denmark, found that postoperative mortality after elective colorectal cancer surgery was significantly lower in patients with high income, higher education, and home ownership compared to home rental. (Frederiksen et al. 2009)

The current maintenance of endorsement submission includes analysis to evaluate the contribution of socioeconomic status (SES) data to the measure. The following SES data were included: race, Hispanic ethnicity, and income. The addition of SES data did not improve measure performance, please see Measure Testing attachment.

McGlynn, E.A., et al., The quality of health care delivered to adults in the United States. N Engl J Med, 2003. 348(26): p. 2635-45.
Calle, E.E., et al., Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med, 2003. 348(17): p. 1625-38.
Benoist, S., et al., Impact of obesity on surgical outcomes after colorectal resection. Am J Surg, 2000. 179(4): p. 275-81.
Breslin, T.M., et al., Hospital factors and racial disparities in mortality after surgery for breast and colon cancer. J Clin Oncol, 2009. 27(24): p. 3945-50.
Morris, A.M., et al., Racial disparities in late survival after rectal cancer surgery. J Am Coll Surg, 2006. 203(6): p. 787-94. Chien, C., et al., Differences in colorectal carcinoma stage and survival by race and ethnicity. Cancer, 2005. 104(3): p. 629-39.
Hardiman, K.M., et al., Disparities in the treatment of colon cancer in octogenarians. Am J Surg, 2009. 197(5): p. 624-8.
Frederiksen, B.L., et al., The impact of socioeconomic factors on 30-day mortality following elective colorectal cancer surgery: a nationwide study. Eur J Cancer, 2009. 45(7): p. 1248-56.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**

McGlynn, E.A., et al., The quality of health care delivered to adults in the United States. N Engl J Med, 2003. 348(26): p. 2635-45.
Calle, E.E., et al., Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. N Engl J Med, 2003. 348(17): p. 1625-38.
Benoist, S., et al., Impact of obesity on surgical outcomes after colorectal resection. Am J Surg, 2000. 179(4): p. 275-81.
Breslin, T.M., et al., Hospital factors and racial disparities in mortality after surgery for breast and colon cancer. J Clin Oncol, 2009. 27(24): p. 3945-50.
Morris, A.M., et al., Racial disparities in late survival after rectal cancer surgery. J Am Coll Surg, 2006. 203(6): p. 787-94. Chien, C., et al., Differences in colorectal carcinoma stage and survival by race and ethnicity. Cancer, 2005. 104(3): p. 629-39.
Hardiman, K.M., et al., Disparities in the treatment of colon cancer in octogenarians. Am J Surg, 2009. 197(5): p. 624-8.
Frederiksen, B.L., et al., The impact of socioeconomic factors on 30-day mortality following elective colorectal cancer surgery: a nationwide study. Eur J Cancer, 2009. 45(7): p. 1248-56.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, Severity of illness, Frequently performed procedure, Leading cause of morbidity/mortality, Patient/societal consequences of poor quality, High resource use

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

Colon and rectal resections constitute approximately 10% of all general surgery procedures performed in the United States

accounting for over half a million procedures performed annually.(Owings et al. 1998) Estimates from the Centers for Medicare & Medicaid Services estimate that colorectal surgery will exceed $77 million in 2009 and over 3,700 hospitals in the U.S. perform colorectal surgery. The most common indications for colon and rectal resections include cancer, diverticular disease, trauma, bowel infarction, and inflammatory bowel disease including Ulcerative Colitis or Crohn's disease. Large bowel cancer is the fourth most common malignancy in the United States, with more than 150,000 new cases in 2009. With 49,920 estimated deaths, it is second only to lung cancer as the leading cause of cancer-related deaths.(Jemal et al. 2009) Upwards of 80% of colon cancer patients are appropriate candidates for curative resection at presentation. Diverticular disease accounts for approximately 130,000 hospitalizations annually in the U.S. and poses significant cost to the health care system.(Munson et al. 1996) Though less than 10% of patients require emergent sigmoid resection for complications of acute diverticulitis,(Stollman et al. 1999) a much higher proportion of patients undergo colon resection under elective settings. With the aging US population, there is an expectation that diverticular disease requiring surgically management as well as patients requiring colorectal surgery for oncologic procedures are on the rise. (Etzione et al. 2009)

The postoperative complication rate for colon and rectal procedures approaches 30% ranking them among the most morbid of all surgical procedures. Failures of adherence to best practices in colorectal surgery are associated with increased complications. (Arriaga et al. 2009) Seventeen percent of patients who undergo colon resections have postoperative ileus and therefore have prolonged hospitalizations (13.8 days as opposed to 8.9 days without ileus, $P < 0.001$). (Iyer et al. 2009) Ten to thirty percent of patients undergoing elective colorectal resections develop surgical site infections.(Prospero et al. 2006, Smith et al. 2004) There are an estimated 600,000 surgical site infections per year for major surgery in the United States, at an estimated cost of $1.8 billion.(Bratzler et al. 2005) The estimated additional costs per surgical site infection in published studies differ widely, from <$400 for superficial infections to $30,000 for serious intra-abdominal infections.(Urban 2006) A study published in 2004 found that approximately half of surgical site infections were detected in the outpatient setting following discharge and accumulated a mean of $6200/patient of home health expenses related to wound care.(Smith et al. 2004) Another review of 1,127 patients undergoing elective colon resections found that those with postoperative deep and organ space infections had a longer hospitalizations as well as markedly higher costs (mean length of stay 21 days versus 6 days and $42,516 versus $10,999, both $P < .001$).(Eagye, Nicolau 2009)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Arriaga, A.F., et al., The Better Colectomy Project: Association of Evidence-Based Best-Practice Adherence Rates to Outcomes in Colorectal Surgery. Ann Surg, 2009.

Bratzler, D.W., et al., Use of antimicrobial prophylaxis for major surgery: baseline results from the National Surgical Infection Prevention Project. Arch Surg, 2005. 140(2): p. 174-82.

Eagye, K.J. and D.P. Nicolau, Deep and organ/space infections in patients undergoing elective colorectal surgery: incidence and impact on hospital length of stay and costs. Am J Surg, 2009. 198(3): p. 359-67.

Etzioni, D.A., et al., Impact of the aging population on the demand for colorectal procedures. Dis Colon Rectum, 2009. 52(4): p. 583-90; discussion 590-1.

Iyer, S., W.B. Saunders, and S. Stemkowski, Economic burden of postoperative ileus associated with colectomy in the United States. J Manag Care Pharm, 2009. 15(6): p. 485-94.

Jemal, A., et al., Cancer statistics, 2009. CA Cancer J Clin, 2009. 59(4): p. 225-49.

Munson, K.D., et al., Diverticulitis. A comprehensive follow-up. Dis Colon Rectum, 1996. 39(3): p. 318-22.

Owings, M.F. and L.J. Kozak, Ambulatory and inpatient procedures in the United States, 1996. Vital Health Stat 13, 1998(139): p. 1-119.

Prospero, E., et al., Surveillance for surgical site infection after hospital discharge: a surgical procedure-specific perspective. Infect Control Hosp Epidemiol, 2006. 27(12): p. 1313-7.

Smith, R.L., et al., Wound infection after elective colorectal resection. Ann Surg, 2004. 239(5): p. 599-605; discussion 605-7.

Stollman, N.H. and J.B. Raskin, Diagnosis and management of diverticular disease of the colon in adults. Ad Hoc Practice Parameters Committee of the American College of Gastroenterology. Am J Gastroenterol, 1999. 94(11): p. 3110-21.

Urban, J.A., Cost analysis of surgical site infections. Surg Infect (Larchmt), 2006. 7 Suppl 1: p. S19-22.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

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

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The current submission for maintenance of endorsement removes venous thromboembolic events (VTE), which include both deep venous thrombosis and pulmonary embolism, from the measure. This change was prompted by recent publications demonstrating that VTE is highly subject to surveillance bias. A study of 2,838 hospitals found that increased VTE prophylaxis adherence was associated with worse risk-adjusted VTE event rates.(Bilimoria, Chung et al. 2013) Paradoxically hospitals with higher quality, identified by number of accreditations and quality initiatives, had worse VTE rates. The explanation for this paradoxical relationship is suggested by the association of higher rates of VTE imaging studies among these hospitals with higher rates of VTE detection. (Bilimoria, Chung et al. 2013, Ju, Chung et al. 2014, Chung, Ju et al. 2015)

Details concerning measure performance with and without inclusion of VTE are included in the Data Testing supplement. The inclusion of socioeconomic status (SES) data has also been evaluated in the Data Testing supplement. Measure performance was evaluated with three additional variables: race, Hispanic ethnicity, and income. These variables did not significantly change or improve the measure performance and, therefore, have not been added to the measure specifications.

Bilimoria, K. Y., J. Chung, M. H. Ju, E. R. Haut, D. J. Bentrem, C. Y. Ko and D. W. Baker (2013). "Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure." Jama 310(14): 1482-1489.
Chung, J. W., M. H. Ju, C. V. Kinnier, M. W. Sohn and K. Y. Bilimoria (2015). "Postoperative venous thromboembolism outcomes measure: analytic exploration of potential misclassification of hospital quality due to surveillance bias." Ann Surg 261(3): 443-444.
Ju, M. H., J. W. Chung, C. V. Kinnier, D. J. Bentrem, D. M. Mahvi, C. Y. Ko and K. Y. Bilimoria (2014). "Association between hospital imaging use and venous thromboembolism events rates based on clinical data." Ann Surg 260(3): 558-564; discussion 564-556.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The outcome of interest is 30-day, hospital-specific risk-adjusted (all cause) mortality, unplanned reoperation, 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). All

outcomes are definitively resolved within 30 days of any ACS NSQIP listed (CPT) surgical procedure. All variables (fields) are explicitly defined in the tradition of the ACS NSQIP and definitions are also submitted in these materials. The original endorsed measure included venous thromboembolism (VTE) as eligible morbidity events, including deep venous thrombosis requiring therapy and pulmonary embolism.

The current set of mortality and major complications for this measure was chosen based on prior work revealing that these complications are related to other important criteria such as large contributions to excess length of stay, large complication burdens, or correlations with mortality. (Merkow et al. 2013) In addition, the desire to limit the outcomes to significant events (ie- some degree of severity according to certain criteria) is the reason that superficial wound infection is excluded from the measure. The current submission removes VTE from the measure as recent publications have demonstrated it is highly subject to surveillance bias. A recent study of 2,838 hospitals found that increased VTE prophylaxis adherence was associated with worse risk-adjusted VTE event rates. (Bilimoria 2013 JAMA) Paradoxically hospitals with higher quality, identified by number of accreditations and quality initiatives, had worse VTE rates. The explanation for this paradoxical relationship is suggested by the association of higher rates of VTE imaging studies among these hospitals with higher rates of VTE detection. (Bilimoria, Chung et al. 2013, Ju, Chung et al. 2014, Chung, Ju et al. 2015)

Bilimoria, K. Y., J. Chung, M. H. Ju, E. R. Haut, D. J. Bentrem, C. Y. Ko and D. W. Baker (2013). "Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure." Jama 310(14): 1482-1489.
Chung, J. W., M. H. Ju, C. V. Kinnier, M. W. Sohn and K. Y. Bilimoria (2015). "Postoperative venous thromboembolism outcomes measure: analytic exploration of potential misclassification of hospital quality due to surveillance bias." Ann Surg 261(3): 443-444.
Ju, M. H., J. W. Chung, C. V. Kinnier, D. J. Bentrem, D. M. Mahvi, C. Y. Ko and K. Y. Bilimoria (2014). "Association between hospital imaging use and venous thromboembolism events rates based on clinical data." Ann Surg 260(3): 558-564; discussion 564-556.
Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the american college of surgeons colectomy composite outcome quality measure. Ann Surg. 2013;257(3):483-489.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Targeted events within 30 days of the index operation are included.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
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.

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).

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.

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 greater than 3 times upper level of the reference range in the setting of suspected myocardial ischemia
c. Physician diagnosis of myocardial infarction.

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.

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 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.

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.

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

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 from minimally contaminated lower respiratory tract (LRT) specimen (e.g. BAL or protected specimen brushing)

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 from preoperative value, but with no requirement for dialysis.

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 five:
a. fever (>38 degrees C),
b. urgency,
c. frequency,
d. dysuria,
e. suprapubic 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
AND ANY ONE or MORE of the following seven:
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.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Patients undergoing any ACS NSQIP listed (primary CPT) colon procedure. (44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44160, 44204, 44205, 44206, 44207, 44208, 44210)

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Cases are collected so as to match ACS NSQIP inclusion and exclusion criteria, thereby permitting valid application of ACS NSQIP model-based risk adjustment. See also exclusions below.

S.10. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*
As noted above, cases are collected so as to match ACS NSQIP inclusion and exclusion criteria, thereby permitting valid application of ACS NSQIP model-based risk adjustment. Therefore, trauma and transplant surgeries are excluded as are surgeries not on the ACS NSQIP CPT list as eligible for selection (see details in next item). Patients who are ASA 6 (brain-death organ donor) are not eligible surgical cases. Of note, the measure excludes patients identified as having had prior surgical procedures within 30 days of a potential index procedure, since this measure is based on 30 day outcomes. A patient who is identified as having had a prior surgical procedure within 30 days of the index case being considered is excluded from accrual. A patient who has a second surgical procedure performed within 30 days after an index procedure has the second procedure recorded as a "Return to the operating room within 30 days" (one of the outcomes defined), but the second procedure cannot be accrued into the program as a new index procedure.

S.11. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
CPT Codes: Procedures not eligible for selection are excluded. (Measure only includes colon procedures, CPTs: 44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44160, 44204, 44205, 44206, 44207, 44208, 44210)
MAJOR TRAUMA: A patient admitted to the hospital with acute trauma and multisystem injury who has surgery for the traumatic injury is excluded.
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, though any operation performed after the patient has been discharged from the transplant stay is eligible for selection. Donor procedures on living donors are not excluded unless meeting other exclusion criteria.
ASA CLASS 6: A patient classified as ASA Class 6 is not eligible for inclusion.

As noted above, the measure excludes patients identified as having had prior surgical procedures within 30 days of a potential index procedure, since this measure is based on 30 day outcomes. A patient who is identified as having had a prior surgical procedure within 30 days of the index case being considered is excluded from accrual. A patient who has a second surgical procedure performed within 30 days after an index procedure has the second procedure recorded as a "Return to the operating room within 30 days" (one of the outcomes defined), but the second procedure cannot be accrued into the program as a new index procedure.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
There is no stratification of this risk-adjusted measure.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
ACS NSQIP performs hospital-level profiling by reporting case-mix adjusted and risk-adjusted postoperative outcomes. The statistical modeling is performed in three steps, which include case-mix adjustment, variable selection, then risk adjustment, all of which are carried out using the SAS software package (v 9.2).

In the first step, clinically similar procedures (defined by CPT codes) are categorized into established groups. Generalized linear mixed modeling (GLMM, also called hierarchical modeling in this measure) is used to calculate linear predictor values for each procedure group (SAS PROC GLIMMIX). These linear predictors (referred to as “CPT Risk”) rank each procedure group on a continuous scale based on the log probability for outcome, and are risk adjusted for patient factors. The CPT Risk variable provides case-mix adjustment for the hospital profiling.

For variable selection of risk factors, step-wise logistic regression (SAS PROC LOGISTIC) is performed using NSQIP predictors. The NSQIP predictors demonstrating statistical significance ($P<0.05$) are selected for the preliminary predictor list. A subset of this list is chosen based on clinical relevance, statistical importance, and ease of data extraction to create a small, fixed or “parsimonious” predictor set (described in: Merkow, Hall et al. 2013) This composite mortality or any serious morbidity outcome measure was evaluated based on the following six predictors: ASA class, CPT risk, functional status, operative indication, emergency case and wound class. Operative indication was categorized into eight separate groups based on ICD-9/ICD-10 codes: cancer, diverticular disease, enteritis/colitis, hemorrhage, volvulus, obstruction/perforation, vascular insufficiency and other.

In the final step, both case-mix adjustment and risk adjustment are performed for the hospital profiling using the CPT Risk and the parsimonious predictor set, respectively. A GLMM is created (SAS PROC GLIMMIX) which reflects the hierarchical nature of the data, with patients clustered within hospitals (random intercept, fixed slope model with logistic regression). The model incorporates the empirical Bayes method, which optimally combines information from the particular hospital with information from the sample of all hospitals to arrive at a best prediction about each hospital’s performance. Sometimes called a reliability adjustment, but more properly described as smoothing or pooling, this adjustment tends to shrink predicted hospital performance towards the grand mean hospital value, with the effect of shrinkage greatest when the hospital sample size is small and when the hospital’s estimate is extreme compared to other hospitals.

Hospital performance is reported as an odds ratios (the odds for the hospital versus the odds for the statistically constructed average hospital). Hospitals with odds ratios less than 1.0 demonstrate better than average performance; those with odds ratios greater than 1.0 demonstrate worse than average performance. Odds ratios are reported with 95% confidence intervals: if the interval does not overlap 1.0, the hospital is designated as a statistically significant high or low outlier, depending on whether the interval is entirely above or below 1.0, respectively.

An outcome was defined as 30-day mortality or any serious morbidity including: cardiac arrest requiring CPR, myocardial infarction, sepsis, septic shock, organ space SSI, deep incisional 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, urinary tract infection, or return to the operating room, according to ACS NSQIP definitions.

Reliability was assessed using a standard method (described in: Huffman, Cohen et al. 2015), which uses information provided by a

random intercept, fixed slope, hierarchical model (implemented by SAS PROC GLIMMIX). Please see Measure Testing attachment.

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
Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the american college of surgeons colectomy composite outcome quality measure. Ann Surg. 2013;257(3):483-489.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
A detailed description of the parsimonious colon surgery outcome measure has been published recently (as described in: Merkow, Hall et al. 2013).

Merkow RP, Hall BL, Cohen ME, et al. Validity and feasibility of the american college of surgeons colectomy composite outcome quality measure. Ann Surg. 2013;257(3):483-489.

S.16. **Type of score:**
Ratio
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
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. See also the risk adjustment methodology section.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
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 (prior estimated sample size for reliability 0.4 was approximately 63 cases), 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 can not be achieved, hospitals may collect data on all eligible patients.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
ACS NSQIP has placed a very high value on accuracy of data collection while maintaining a sample size large enough for statistical modeling and keeping within regulations for patient confidentiality. The methodology of our program has been highly successful with increasing numbers of participants every year, and measureable improvements in surgical outcomes over time based on the

O/E ratios for mortality and various post-surgical complications. Historically, the use of trained data collectors within ACS NSQIP and a comprehensive support system has resulted in high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Data definitions are continually evaluated and inter-rater reliability audits are regularly performed.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Registry, Management Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Model is based on historical ACS NSQIP Data file. Data sources are as above- collection is consistent with historical ACS NSQIP approaches to data collection. Model is based on ACS NSQIP but measure would not require participation in ACS NSQIP.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
URL

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0706_MeasureTesting_Maintenance-May2016.doc

NATIONAL QUALITY FORUM

NQF #: 0706 NQF Project: Patient Outcomes Measures: Phases I and II

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)
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

field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

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. Data sample for hospitals would be one year sampling according to systematic algorithm and sample and reliability information supplied: e.g. reliability of 0.4 would require roughly 63 cases/annum.

May 31, 2016 Maintenance of Endorsement Update:
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 108,571 patient records for eligible colon surgeries from 504 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 = 451, cases = 33,747), 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 for hierarchical risk adjustment model, incorporating procedure risk score. 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	INTEPRETATION
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 0.0106. Using a minimum acceptable reliability for mortality/serious morbidity in patients =65 of 0.4, the proportions of hospitals likely to have a “minimally acceptable” reliability estimate are as follows. 42.9% of all U.S. hospitals and 68.7% of ACS NSQIP hospitals meet the 0.4 reliability requirement. These ~40% of US hospitals perform roughly 85% of all colectomies in the country. This level of reliability is comparable to or exceeds published figures for other approved measures- the ACS provides this reliability data on all submitted measures despite the fact that many measure developers do not submit comparable data.

Furthermore, it is also expected that as the population and diversity of institutions participating in this measure increases, the reliability will increase as well- making our initial estimate a conservative one. [This is related to the ACS NSQIP having some bias toward larger academic institutions.] Furthermore, we do provide in our results below information on increasing the reliability by increasing the sample size, which would be considered for any implementation. However, there will always be a trade-off between drafting a measure with higher reliability but having it apply to fewer institutions (since requiring increasing sample size will exclude more and more institutions).

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	41	55.8	79.6
0.4	63	42.9	68.7
0.5	94	31.6	53.5
0.6	141	20.1	27.5
0.7	219	9.6	3.8

* 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 451 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 colon surgery outcomes measure: (1) dropping venous thromboembolism (VTE) 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	VTE+, SES-	VTE+, SES+	VTE-, SES-	VTE-, SES+
0.00-0.20	35.03	35.92	33.92	35.70
0.21-0.40	33.48	33.92	33.70	32.82
0.41-0.60	24.39	23.73	24.17	24.17
0.61-0.80	7.10	6.43	8.20	7.32

VTE+, SES- = VTE included, SES not included (original model)
VTE+, SES+ = VTE included, SES included
VTE-, SES- = VTE not included, SES not included
VTE-, SES+ = VTE not included, SES included

Using a minimum acceptable reliability of 0.4 for the colon surgery outcomes measure, the proportions of hospitals with a "minimally acceptable" reliability estimate is slightly above 30% across all four variations.

The mean number cases per hospital in the 2014 data set was 74.8, but there was positive skew in the distribution of sample sizes (median=51). 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.

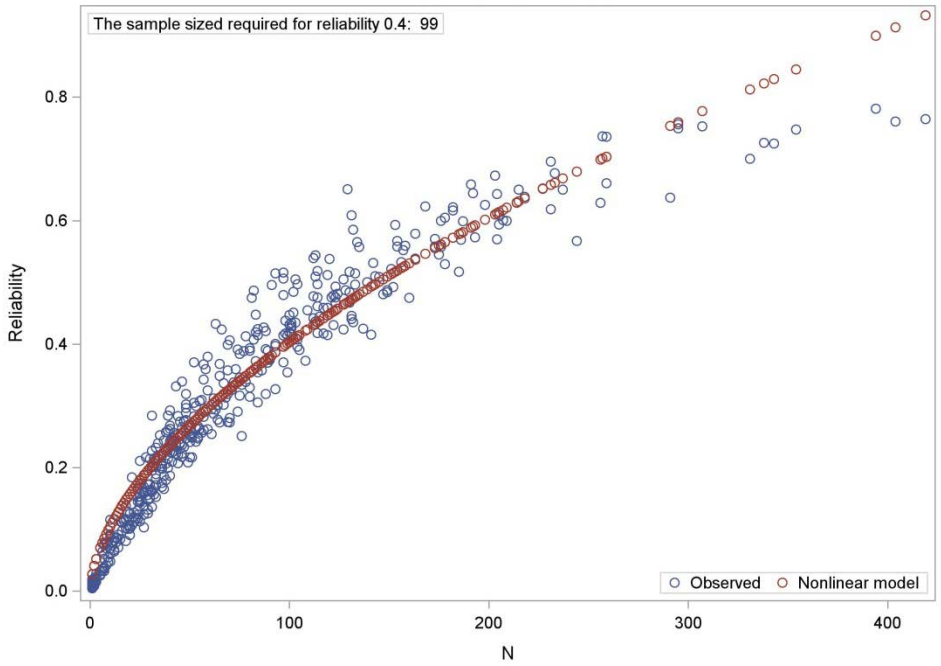


Figure: Reliability vs Case number for colon surgery outcomes measure (without VTE and without SES), regression and observed

Thus, an empirical estimate of the sample size required to achieve reliability of 0.4 is 99. This number would be considered an appropriate and achievable target for the number of colon cases for hospitals interested in participating in this measure. Nevertheless, hospitals with fewer cases might still benefit from participation in this measure

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

The importance of tracking morbidity and mortality after colon surgery is emphasized in the evidence and consistent with the current measure specifications. Please see criterion 1c for additional information.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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:
Models were constructed using a large sample derived from the ACS NSQIP database for 2011-2014. See 2a1 above.

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:

See Risk-adjustment Methodology in Specifications. C-statistics and Hosmer-Lemeshow P-values for the developmental data set were computed; c-statistics were computed for an independent validation data set base on 2007 data.

May 31, 2016 Maintenance of Endorsement Update:

See Risk-adjustment Methodology in Specifications.

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), and for with and without SES variables, in order to assess their effects on model quality with respect to discrimination and calibration.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
See Risk-adjustment Methodology in Specifications. Model validity (a similar c-statistic, discrimination) was demonstrated when the 2008 model was applied to 2007 data.

May 31, 2016 Maintenance of Endorsement Update:

See Risk-adjustment Methodology in Specifications.

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 VTE as an outcome and the presence versus absence of SES in the prediction equation.

Model	Model c statistic	2010 c statistic	Model HL	p-value	2010 HL	2010 p-value	Model Brier	2010 Brier
Colon measure With VTE, Without SES	0.7177	0.7241	30.2925	0.0002	26.3604	0.0009	0.1296	0.1421
Colon measure With VTE, With SES	0.7183	0.7245	26.2982	0.0009	21.8053	0.0053	0.1295	0.1421
Colon measure Without VTE, Without SES	0.7166	0.7209	38.2492	0.0000	27.3687	0.0006	0.1232	0.1373
Colon measure Without VTE, With SES	0.7174	0.7216	32.9696	0.0001	26.1840	0.0010	0.1232	0.1372

2011	C Statistic	HL	p_value	Brier
Colon measure with VTE, without SES	0.7086	28.6484	0.0004	0.1364
Colon measure with VTE, with SES	0.7096	24.7120	0.0017	0.1363
Colon measure without VTE, without SES	0.7073	32.0542	0.0001	0.1310
Colon measure without VTE, with SES	0.7085	32.5091	0.0001	0.1309

2012	C Statistic	HL	p_value	Brier
Colon measure with VTE, without SES	0.7192	18.8144	0.0159	0.1277
Colon measure with VTE, with SES	0.7199	12.3570	0.1360	0.1277
Colon measure without VTE, without SES	0.7181	21.6632	0.0056	0.1218
Colon measure without VTE, with SES	0.7190	14.3001	0.0743	0.1217

2013	C Statistic	HL	p_value	Brier
Colon measure with VTE, without SES	0.7130	21.7377	0.0054	0.1308
Colon measure with VTE, with SES	0.7139	25.3037	0.0014	0.1308
Colon measure without VTE, without SES	0.7131	21.6032	0.0057	0.1239
Colon measure without VTE, with SES	0.7141	26.6418	0.0008	0.1239

Jul 1, 2011 - Jun 30, 2012	C Statistic	HL	p_value	Brier
Colon measure with VTE, without SES	0.7260	10.6689	0.2212	0.1256
Colon measure with VTE, with SES	0.7262	7.5702	0.4765	0.1256
Colon measure without VTE, without SES	0.7241	11.6855	0.1658	0.1190

Colon measure without VTE, with SES	0.7243	10.0459	0.2618	0.1190

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(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 2011-2014. See 2a1 above.

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
See Risk-adjustment Methodology in Specifications.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
See Risk-adjustment Methodology in Specifications.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 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)*:
The data sample is derived from the most recent ACS NSQIP Data file (2008). The Colorectal model used 21,694 patient records. Future models can be constructed using the most recent Data file. If this measure is adopted by sufficient numbers of non-NSQIP hospitals re-modeling can be based on data from the broader sample of hospitals.

May 31, 2016 Maintenance of Endorsement Update:

The data sample of 108,571 colon cases included in ACS NSQIP Data files (2011-2014) were used for model generation.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
Preliminary risk-adjustment models were constructed for these developmental purposes using step-wise logistic regression. Compared to hierarchical models this methodology poses fewer convergence problems, has step-wise variable-selection methodology, and we have found that it provides nearly identical risk-adjustment as random intercept hierarchical models. Odds ratios and parameters reported here are derived from hierarchical model methodology applied to the predictor set established using step-wise logistic regression methods. See all other details on risk adjustment described elsewhere above, including generation of CPT risk score, above (Measure specifications- risk adjustment methodology) and following (2e3).

May 31, 2016 Maintenance of Endorsement Update:

It was our intention to estimate model parameter values (more accurately) from a large, multi-year sample and then apply historical prediction equations to samples composed of annual accumulations of data. This required a logistic rather than a hierarchical approach (which involves contemporaneous data modeling), with the quality metric being an O/E ratio. Step-wise logistic regression, informed by clinical insights and the need for parsimony, resulted in prediction equations with either 6 factors (CPT category, ASA Class, Functional Status, Indication, Emergent, Wound Class) or 9 (with 3 additional variables for exploration of SES: Median Income, Hispanic Ethnicity, Race).

We examined differences in risk-adjusted outcomes when equations were applied to the 2014 data. Specifically, when the SES variables were or were not included, and when VTE was or was not included as an outcome (thus, there were 4 sets of parameter values). SES was examined to determine whether this factor would represent a crucial risk-adjustment component, and VTE was examined as there is evidence that the observation of a VTE event is subject to a substantial surveillance bias such that it is an inappropriate outcome for quality monitoring – "good" hospitals that initiate careful, sometimes universal, surveillance are at a disadvantage compared to other hospitals that are less likely to look for non-symptomatic VTEs.

Hierarchical modeling provides several theoretical advantages over ordinary logistic modeling including appropriate consideration of the nested structure of data (patients within hospitals) and the automatic incorporation of an empirical-Bayes-type shrinkage adjustment to stabilize estimates (of particular importance when sample sizes are small and event rates low). Our own research has indicated the adjustment of error variance estimates associated with nesting has little practical effect. However, shrinkage adjustments do provide, under certain conditions, for better quality estimates, although shrinkage can potentially mask real quality differences (i.e., the approach can be overly conservative). While not reported on in this submission, we are exploring the incorporation of post-logistic modeling smoothing (shrinkage) to measure colon surgery O/E ratios. This methodology, as applied generally to ACS NSQP data, has been described elsewhere (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*, in press.)

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
See Risk-adjustment Methodology in Specifications. 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 20 predictors for inclusion in the model. In order of inclusion these variables were: ASA Class, pre-operative Functional Status, Indication, (Log Odds CPT) "CPT Risk", Emergent, Wound Class, Dyspnea, Weight Loss, Steroid Use, Smoking, Disseminated Cancer, History of COPD, Ascites, Hypertension, Ventilator Dependent, Age Group, Radio Therapy, Alcohol Use, Bleeding Disorder, and Previous Vascular Event/Disease. The c-statistic was 0.738 and the Hosmer-Lemeshow was 0.043. 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 six selected variables (ASA Class, pre-operative Functional Status, Indication, (Log Odds CPT) "CPT Risk", Emergent, and Wound Class), the c-statistic was 0.727 and the Hosmer-Lemeshow was 0.177). The use of these six predictors for modeling was further evaluated. Using a 95% confidence interval for the ratio of observed to expected events (O/E), this six-variable logistic model identified 16 statistical outliers (10 low outliers and 6 high outliers). When the same six 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), 17 outliers were detected (11 low outliers and 6 high outliers). Thus, using a 95% confidence interval, logistic and hierarchical models identified 3% of hospitals as high outliers. See additional data on reliability and sample size estimation provided above (Scientific Acceptability-reliability testing).

May 31, 2016 Maintenance of Endorsement Update:
Using a 95% confidence interval for the observed to expected events (O/E) ratio, the original colon surgery outcome measure (without SES and with VTE) identified 12 low and 10 high outliers among the 451 hospitals with data in 2014. The addition of SES data resulted no changes in outlier status among these hospitals (weighted kappa = 1). In addition, 22 hospitals increased decile status by 1 category and 22 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 colon surgery outcomes measure.

We also examined the effect of removing VTE for models without SES variables. The comparison of outlier determinations is shown below (without-VTE in the column, and with VTE in the row, weighted kappa = 0.765)

Outlier Status	Low	None	High
Low	10	4	0
None	2	424	3
High	0	1	7

Changes in decile status are shown below

Differences:	Change based on exclusion of VTE (Decile Without VTE - Decile With VTE)
-6	2

-5	0
-4	2
-3	6
-2	7
-1	62
0	258
1	114

Thus, the inclusion versus exclusion of VTE has important effects on outlier and decile status.

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment: Risk adjusted

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe 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 Strategy Data Sample Section (2b4.1).

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:

The default methodology for discrimination performance is based on the computed 95% CI (using Ulm's method) for the O/E ratio. If the interval is entirely above1.0, the hospital is identified as having performance significantly worse than expected. If the interval is entirely below 1.0, the hospital is identified as having performance significantly better than expected. If the interval overlaps 1.0 the hospital is performing "as expected." Depending on programmatic objectives, the implementing organization could also opt for outlier status being defined by upper and lower percentile ranks in O/E ratios.

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
See Risk-adjustment strategy Testing Results (2b4.3)

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

May 31, 2016 Maintenance of Endorsement Update:

The only sources of data are those indicated above. This measure requires clinical data (electronic or paper records), with administrative data (zip code) added only as necessary. The current maintenance of endorsement submission provides measure performance with the addition of the following SES data: race, Hispanic ethnicity, and income, as estimated by proxy using patient zip code mapped via the University of Michigan Population Studies Center zip code characteristics (available at http://www.psc.isr.umich.edu/dis/census/Features/tract2zip/)

The advantage of clinical data versus administrative or claims data in identifying risk-adjusted outcomes is exemplified in the study by Steinberg et al (2008). The study compared comorbidities collected and postsurgical complications from the ACS NSQIP database and the University HealthSystem Consortium (UHC). Comorbidities per patient were identified twice as often in the UHC system, while there was a discordance of 26% in identifying complications (UHC complication rate, 2% vs. ACS NSQIP complication rate, 28%). Recent studies have compared ACS NSQIP data and Medicare claims data, indicating lack of agreement and poor correlation between the two data sources as it relates to complication identification and risk-adjustment.(Lawson, Zingmond et al. 2015, Lawson, Louie et al. 2016) Using administrative or claims data may result in significant differences in risk-adjusted outcomes than using clinical data.

Lawson, E. H., R. Louie, D. S. Zingmond, G. D. Sacks, R. H. Brook, B. L. Hall and C. Y. Ko (2016). "Using Both Clinical Registry and Administrative Claims Data to Measure Risk-adjusted Surgical Outcomes." Ann Surg 263(1): 50-57.
Lawson, E. H., D. S. Zingmond, B. L. Hall, R. Louie, R. H. Brook and C. Y. Ko (2015). "Comparison between clinical registry and medicare claims data on the classification of hospital quality of surgical care." Ann Surg 261(2): 290-296.
Steinberg, S.M., et al., Comparison of risk adjustment methodologies in surgical quality improvement. Surgery, 2008. 144(4): p. 662-7; discussion 662-7.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
See above

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
See above

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*:

Measure is not currently stratified; measure is case mix adjusted. As mentioned above, the current submission includes measure testing with SES data, including race, ethnicity and income (estimated using zip code, as described above). Please see Testing Results (in particular 2a2.3, 2b2.3 and 2b4.3) for additional details regarding model performance when SES data is included.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
N/A

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Data generated as byproduct of care processes during care delivery (Data are generated and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition), Coding/abstraction performed by someone other than person obtaining original information (E.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or registry), Other
If other: dedicated abstraction personnel

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed*

to compute the performance measure score are in defined, computer-readable fields)
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
A completely electronic medical record (EMR) would be needed to capture all risk factors that enter into the model. In addition, a software module (currently available to ACS NSQIP subscribers) will be required to transfer information from the EMR to a measure submission database. The ACS NSQIP is in the process of developing an automated process with EMR vendors, however, electronic entry for this measure is not currently available.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
ACS NSQIP has been open to subscription by private sector hospitals since 2004. Ten years prior to this time the program was implemented in the U.S. Department of Veterans Affairs. Thus we have long term experience with the data collection and operational use of the O/E ratio for quality improvement and benchmarking on which this measure is based. Historically, the use of trained data collectors within ACS NSQIP and a comprehensive support system has resulted in high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Data definitions are continually evaluated and inter-rater reliability audits are regularly performed.

ACS NSQIP has placed a very high value on accuracy of data collection while maintaining a sample size large enough for statistical modeling and keeping within regulations for patient confidentiality. The methodology of our program has been highly successful with increasing numbers of participants every year, and measureable improvements in surgical outcomes over time based on the O/E ratios for mortality and various post-surgical complications. Due to the much smaller number of variables needed for participation in this measure than in the full program, we expect that hospitals that are not ACS NSQIP participants will also be able to achieve highly reliable results.

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Public Reporting https://www.medicare.gov/hospitalcompare/acs-surgical-measures.html Hospital Compare Quality Improvement with Benchmarking (external benchmarking to multiple organizations) ACS NSQIP https://www.facs.org/quality-programs/acs-nsqip Quality Improvement (Internal to the specific organization) ACS NSQIP https://www.facs.org/quality-programs/acs-nsqip

4a.1. **For each CURRENT use, checked above, provide:**

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

Public Reporting: According to the data reporting period for April 2016, there are 131 hospitals currently reporting their risk-adjusted surgery outcomes data for NQF-endorsed measures from ACS. Please see https://www.medicare.gov/hospitalcompare/acs-surgical-measures.html for more information. (Accessed 5/16/2016)

Quality Improvement, both internal and external with benchmarking: There are over 600 hospitals currently participating in ACS NSQIP and receiving risk-adjusted benchmarking reports. ACS NSQIP hospitals utilize their internal data for the purpose of quality improvement initiatives specific to the organization.

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

A recent analysis indicates that over 8 years in the program, 62% and 71% of hospitals improve their performance in mortality and risk-adjusted complications. (Cohen et al. 2016) Annual reductions are approximately 0.8% in mortality and 3.1% in morbidity; though small, these reductions provide cumulative benefit as hospitals continue participation in the ACS NSQIP program. For 2014, there were 451 hospitals contributing 33,747 colorectal surgery cases. The O/E ratios for colon surgery mortality and serious morbidity range from 0.66 to 1.4 for participating hospitals. These numbers indicate that although there have been improvements

over time, a performance gap remains between those performing better and worse than expected after risk and case mix adjustment. The interquartile range for the O/E ratio is 0.15, and the 10th percentile and 90th percentile O/E ratios were 0.86 and 1.17, respectively.

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

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

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.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB).

0697 : Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure

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

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications completely harmonized?

Yes

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
NA - different target populations

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American College of Surgeons
Co.2 **Point of Contact:** Sameera, Ali, sali@facs.org, 312-202-5431-
Co.3 **Measure Developer if different from Measure Steward:** American College of Surgeons
Co.4 **Point of Contact:** Sameera, Ali, sali@facs.org, 312-202-5431-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Clifford Ko
Sameera Ali
Bruce Hall
Mark Cohen
Yaoming Liu
Julia Berian

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.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2011
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0713
Measure Title: Ventriculoperitoneal (VP) shunt malfunction rate in children
Measure Steward: Boston Children's Hospital, Center for Patient Safety and Quality Research
Brief Description of Measure: This measure is a 30-day malfunction rate for hospitals that perform cerebrospinal ventriculoperitoneal shunt operations in children between the ages of 0 and 18 years.
Developer Rationale: Ventricular shunt malfunction places children at risk for potentially irreversible neurologic system deficits and death if not treated promptly. Shunt malfunction treatment is associated with the need for hospitalization and re-operation. The hospitalization itself is disruptive to the child and family, which may lead to impaired quality of life. The need for re-operation places the child at additional risk for central nervous system infection and other adverse events.

Numerator Statement: The number of initial ventriculoperitoneal (VP) shunt placement procedures performed on children between the ages of 0 and 18 years of age that malfunction and result in shunt revision within 30 days of initial placement.
Denominator Statement: The total number of initial cerebrospinal VP shunt procedures performed on children between the ages of 0 and 18 years.
Denominator Exclusions: Patients with evidence of VP shunt placement or removal in the year prior to their index procedure are excluded.

Measure Type: Outcome
Data Source: Electronic Clinical Data
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2011 **Most Recent Endorsement Date:** Sep 20, 2012

Maintenance of Endorsement – Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This maintenance measure, initially endorsed in 2011, calculates ventriculoperitoneal (VP) shunt malfunction rates in children between 0 and 18 years. Clinical action(s) to impact the outcome is identified as treating the malfunction promptly.
- The developer submitted new evidence on the revision rate for VP shunts and prediction of shunt failure, as well as shunt malfunction risk factors.
- As a rationale for measuring this health outcome, the developer notes that VP malfunction places children at risk for irreversible neurological damage if not treated promptly. VP shunt malfunction treatment is associated with the need for additional hospitalization which is disruptive to the child and family. VP malfunction treatment is also associated with additional operations which places the child at risk for further damage to the nervous system.

Question for the Committee:

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

Guidance from the Evidence Algorithm: Health outcome (Box 1) →relationship between outcome and at least one healthcare action identified/supported by stated rationale (Box 2) → Pass

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

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

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

- Data from the most recent 3 year reporting period (CY12 Q2 – CY15 Q1) show that of 46 eligible cases at Boston Children's Hospital, 2 (4.35%) experienced a malfunction within 30 days. The shunt malfunction rate (SMR) was 1.42 [95% CI (0.16, 5.11)], which was not significantly different from the null value of 1.

Disparities

- At the time of initial (time-limited) endorsement in 2011, the developer noted that it was testing disparities to inform results stratification presentation by race/ethnicity.
- With the current submission, the developer notes that disparities in shunt malfunction have not been assessed; however, the developer cites two studies examining shunt failures. *NQF guidance states that by the time of maintenance of endorsement evaluation, the measure should be in use and data from implementation of the measure should be submitted rather than using data from literature.*
  - A retrospective cohort study of 466 patients and 739 operations (with failure rates at 24.1% after 90 days and 29.9% after 180 days), where no demographic, clinical or procedural data predicted shunt failure.
  - A retrospective, longitudinal cohort study of 1,307 patients from 32 hospitals found that patients in the Midwest were more likely to experience multiple shunt revisions. Lower revision rates were associated with higher hospital volume of initial shunt placement. There was wide variation in rates of VP shunt revision among children's hospitals.

Questions for the Committee:

- *Does data from a single hospital provide sufficient information to assess whether there is a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1a.

- Esoteric - Maybe more suitable as a non-PQRS QCDR measure?
- Not collected in any national registry
- Only 9/12 malfunctions identified by ICD9 code
- Very low complication rate -- unable to discern difference from 0
- Is this for facility or personal reporting?"
- "This is an outcome measure submitted for maintenance. The paper by Rossi et al referenced under 1b 5 is actually a single institution retrospective review that was unable to identify risk factors for shunt malfunction in children. The paper concluded that the causes of shunt malfunction was not known and ""beyond surgeon influence"" and therefore shunt malfunction rate should be questioned as a quality measure.
- The staff asked, ""Is there at least one thing that the provider can do to achieve a change in the measure results?"" While there is not a large body of objective evidence identifying what specific processes are linked to outcome, it is widely accepted in the pediatric neurosurgical community that VPS malfunction rate is modifiable and is a realistic measure of overall process of care within an institution. This reviewer cannot state that there is new evidence in support of this, the measure is submitted for maintenance and I am not aware of evidence that the link between this outcome measure and processes of career is less supportable than at the time of initial endorsement and therefore I recommend a PASS.

1b.

- This reviewer has difficulty understanding the paucity of performance data reported on a measure that has been approved since 2011. The developers report a three year rolling average from their own institutions (BCH) and suggest this is close to a benchmark derived from 10 hospitals in the PHIS database (an administrative database of the Children's Hospital Association). I do not appreciate that the developers used the PHIS database to determine whether or not significant shunt malfunction rate differences occurred among participating PHIS institutions. I believe this could have been readily accomplished and do not see an explanation as to why this was not done.
- The answer to the question on SDS disparities is, in my opinion insufficient. The authors state ""these data are not required to be collected at the institutions utilizing the measure..."" I do not understand these response. SDS data are readily available A large proportion of patients requiring VPS placement are premature infants with intracranial bleeds. SDS factors have been found to be important predictors of outcome in a number of other health outcomes in this population. In my opinion, this is a correctable deficiency in this measure. I agree with the preliminary rating by the staff of INSUFFICIENT in this category.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Data for this measure, as specified, is electronic clinical data from the Pediatric Health Information System (PHIS), a database comprised of 42 tertiary care pediatric hospitals in the US.
- The developer states that data from CY12 Q2 – CY15 Q1 were obtained from Boston Children's Hospital's electronic health record and administrative claims database.

Specifications:

- The measure is specified as a facility-level measure for the hospital/acute care setting.
- The denominator includes the total number of initial cerebrospinal VP shunt procedures performed on children between the ages of 0 and 18 years.
- The numerator includes the number of VP shunt placement procedures in the denominator that malfunction and result in shunt revision within 30 days of initial placement.
- ICD-9 and ICD-10 codes, and a conversion table are provided in S.2b. The developer states that formal analysis of reliability has not been performed due to reliance on ICD9-ICD-10 codes to calculate the measure.
- Exclusions, identified since initial review, are patients with evidence of VP shunt placement or removal in the year prior to their index procedure. Exclusions details are shown in S.11.

- This outcome measure is risk-adjusted using a statistical risk-adjustment model with 6 factors: age of insertion, congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity, and spina bifida. However, it is unclear how risk-adjustment is incorporated into the measure, if calculated as described in section S.18.
- In the most recent three-year reporting period (CY12 Q2 – CY15 Q1), there were 46 eligible cases of VP shunt placement. Two (4.35%) experienced malfunction within 30 days.
- Due to the small number of VP procedures, the measure is reported as a three year rolling rate.

Questions for the Committee :

- o *Are all the data elements clearly defined? Are all appropriate codes included?*
- o *Is the logic or calculation algorithm clear?*
- o *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Summary of reliability testing from the prior review:

- In the initial consideration of this measure, the Committee expressed concern that the measure had limited testing data from a single institution but agreed to recommend it for time limited endorsement because the measure was important "to measure and report as an outcome because it addresses a high-impact procedure" for pediatric patients.
- Testing data submitted that included data from the period CY06-CY09 Q3 are provided with analyses for CY08 and CY09Q3. The data is displayed for Children's Hospital of Boston (CHB) and other PHIS participating hospitals (benchmark).

Table 1: Three year rolling 30-day VP Malfunction Rates, CY08 & CY09Q3

	Procedures (N)	Complications (N)	Malfunction Rate (%)	95% Confidence Interval
CY08				
CHB	44	2	4.6	0.6, 15.5
Benchmark	3351	294	8.8	7.8, 9.8
CY09Q3				
CHB	42	3	7.1	1.5, 19.5
Benchmark	3366	300	8.9	8.0, 9.9

Benchmark: CHB and all PHIS hospitals combined

Describe any updates to testing:

- Reliability testing was not conducted. Data element validity testing was done. NQF guidance provides that if data element validity testing is done, additional data element reliability testing is not required.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Data element testing was performed. In 2010, medical records of 107 patients who had had shunt procedures in two children's hospitals were evaluated against the ICD-9 coding algorithm.
- Of the 107 medical records, 52 (100%) of the cases represented at Boston Children's and 60 cases (a 25% sample that included 5 with coded shunt malfunction and 55 randomly selected procedures with no malfunction codes) from the second children's hospital were evaluated against the ICD-9 coding algorithm for shunt malfunction.
- An expert panel was convened to verify accuracy of ICD-9 to ICD-10 conversion, accomplished using a 3M ICD-10 Code Translation tool. The panel concluded that the numerator and denominator were appropriately identified whether using ICD-9 or ICD-10 codes.

Results of reliability testing

- The ICD-9 coding algorithm correctly identified 9 of 12 malfunctions found by chart review with a sensitivity of 0.75. Specificity was high at 0.96 and a false negative rate of 3%.
- *Sensitivity* measures the proportion of actual positives that are correctly identified as such. A sensitivity value of 0.75 reflects the accuracy of the algorithm in identifying a shunt malfunction when present in the medical record (the authoritative source).
- *Specificity* measures the proportion of actual negatives that are correctly identified as such. A specificity value of 0.96 reflects the accuracy of the algorithm in documenting the absence of shunt malfunction when it is not recorded in the medical record (the authoritative source).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *What does the Committee expect with regard to understanding and evaluation of missing data?*

Guidance from the Reliability Algorithm:

Precise specifications (Box 1) → No empiric reliability testing conducted (Box 2) →Empirical validity testing of the data elements conducted (Box 3) →skip to Validity algorithm (Box 10) → Appropriate method of testing (Box 11) →moderate certainty of validity (Box 12a)

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Summarize the validity testing from the prior review:

- In the previous consideration of this measure, the Committee expressed concern that the measure had limited testing data from a single institution. The Committee agreed to recommend it for time limited endorsement because the measure was important "to measure and report as an outcome because it addresses a high-impact procedure" for pediatric patients.
- Testing data from CY12 Q4 through CY15 Q1 were submitted in and these data are presented below.

Describe any updates to validity testing:

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Validity testing method:

- Data element testing was performed. In 2010, medical records of 107 patients who had had shunt procedures in two children's hospitals were evaluated against the ICD-9 coding algorithm.
- Of the 107, 52 cases (100%) at Boston Children's and 60 cases (a 25% sample that included 5 with coded shunt malfunction and 55 randomly selected procedures with no malfunction codes) from the second hospital were evaluated against the ICD-9 coding algorithm for shunt malfunction.

Validity testing results:

- The ICD-9 coding algorithm correctly identified 9 of 12 malfunctions found by chart review with a sensitivity of 0.75. Specificity was high at 0.96 and a false negative rate of 3%.
-

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The developer notes there were no exclusions although patients with evidence of VP shunt placement or removal in the year prior to their index procedure are excluded. An analysis of the exclusion was not provided.

2b4. Risk adjustment: **Risk-adjustment method** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

The developer notes that variation in shunt malfunction rates has not been assessed across populations groups due to the fact that these data are not required to be collected at the institutions using the measure.

Risk adjustment summary

- This measure was risk adjusted after univariate and logistic regression analysis of a number of variables suggestive of higher risk of shunt malfunction. The final model adjusted for 6 factors that previous published material had shown affected malfunction rates,.
- A logistic regression model adjusts for age at insertion (0-30 days, 31-365 days, and >1y), congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity, and spina bifida.
- Model performance was assessed using the c-statistic to determine how well the risk adjustment model distinguishes events from non-events. The c-statistic is reported as 0.576. A value of 0.5 indicates that the model is no better than chance at making a prediction of patients with and without the outcome of interest.
- The developer did not identify all of the variables that were tested or how the 6 factors were selected for the final model.
- It is unclear how the predicted (expected) value from the risk-adjustment model is used in the calculation of the measure.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included and used in the measure?*
- *Given the c-statistic result, what recommendations does the Committee have for the developer regarding measure validity?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports that meaningful differences between CHB and other PHIS participating hospitals will be assessed at 3-year intervals. Information provided indicates that this assessment will use 10 specifically identified hospitals, including CHB.
- Note that benchmark institutions were chosen for having a pediatric neurosurgery practice similar to BCH in that they are also academic institutions with similar case volume and distribution.
- The developer reports that benchmark for each year is the mean VP malfunction rate of all participating pediatric hospitals in the Pediatric Health Information System (PHIS) dataset.
- From CY12 Q2 - CY 2015 Q1, BCH's SMR was 1.42, which was not significantly different from the null value of 1.

Question for the Committee:

o *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- Missing data analysis not done.

Guidance from the Validity Algorithm:

Precise specifications (Box 1) → Threats to validity only somewhat assessed (no information on exclusions and missing data and little on meaningful differences) (Box 2) → Insufficient

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

Committee pre-evaluation comments

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

2a.

- The authors report a through approach to converting from ICD 9 to ICD 10. These are well defined parameters and I do not have any concerns with respect to this process or the correctness of codes to identify relevant cases

2a.2

- The data source utilized is an administrative database. This is a source type in which coding errors commonly occur. The authors tested the accuracy of coding versus clinical chart review and found a sensitivity of 75%. This is not acceptable in this reviewers opinion. Further, the reliability testing was only done at one institution. The sensitivity may be even lower at institutions not focusing attention on this issue in a manner comparable to the measure developing institution. The specificity of 96% is acceptable but only assessed at one institution. The staff rated this category as moderate. I would rate it as LOW bordering on INSUFFICIENT.

2b.1

- The specifications seem consistent with the evidence.

2b.2

- Validity was tested at only one institution. This is less than adequate. This measure was endorsed in a time limited fashion by NQF due to clinical importance. There has been ample opportunity for wider validity testing and this has not been completed.
- I have spoken with several leaders in Pediatric Neurosurgery and receive the impression that shunt malfunction rate is widely believed not to be a reliable measure of quality of care. Further using shunt revision as a surrogate for shunt malfunction may not be valid. The indications for revision vary widely between institutions. Therefore, the rate of revision may reflect more about the institutional approach and philosophy or shunt care than about quality of care.

2b.3

- The risk adjustment model has never been adequately developed. By the developers admission, they were not able to identify relevant risk factors that proved to be predictive at a statistically significant level when subjected to logistic regression. I do not see evidence of further development over the 5 years since initial endorsement. The authors reference a study (Rossi et al) in a different portion of the application that suggests that shunt

malfunction rate is not related to controllable surgeon factors and is not an acceptable measure of quality of care. I would agree with the staff preliminary assessment of INSUFFICIENT in this category.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data are collected during the provision of care, via coding and abstraction, and through electronic medical records. No modifications have been made to the data collection method.
- All data elements are in defined fields in a combination of electronic sources.
- There are no fees associated with use of the measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery and are data elements likely to be available in electronic medical records in all settings potentially interested in use?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

I agree with the staff that the variables are collected during care delivery and that the feasibility is HIGH.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure

- The measure is currently used in the Pediatric National Surgical Quality Improvement Project.
- The developer's institution is included in the Pediatric Health Information System (PHIS) which is an administrative database for inpatient, emergency and ambulatory surgery data from 42 institutions. From CY 2012 second quarter through CY 2015 first quarter, there were 1, 121 shunts and 37 malfunction cases.
- The measure is used for internal quality improvement purposes and is included in a summary of hospital-wide initiatives.

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- The Pediatric National Surgical Quality Improvement Project (P-NSQIP) is a result of a collaboration between the American College of Surgeons and the American Pediatric Surgical Association. P-NSQIP is open to all pediatric hospitals, freestanding general acute care children's hospitals, children's hospitals within a larger hospital, specialty children's hospitals, or general acute care facilities with a pediatric wing.

Improvement results

- The developer reports that for the past 12 quarters (CY 2011, fourth quarter through CY 2014, third quarter), VP malfunction rate has remained close to the benchmark of 1. At BCH, there were 3 eligible cases of shunt malfunction in 2011 and 2012. There have been no eligible cases since December 2012 through the third

quarter of CY 2014.

Unexpected findings (positive or negative) during implementation:

- The developer states there were no unintended consequences found during testing.

Potential harms

- No potential harms identified.

***Questions for the Committee**:*

- *Given the reported use of the measure to date, what would the Committee like to see to increase value of measure use?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*
- *What steps have developers taken to address stratification in order to see changes in performance?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- The authors state that the measure is used by Pediatric NSQIP. This is part of a new pilot initiative in Pediatric NSQIP. The NSQIP collects many outcome variables and is a program designed for intra and inter institutional use for collaborative quality improvement. Use of a variable for inclusion in NSQIP indicates interest in the field in measuring the variable but does not imply that it is universally accepted as a measure of quality that meets standards analogous to NQF endorsement. I am unaware the measure is used for public reporting either within NSQIP or outside. NSQIP does not support or advocate public reporting. I am not aware of widespread or even moderate use of the measure. The leading national quality consortium on VP shunt quality of care is the Hydrocephalus Research Network (HCRN). I am not aware that the HCRN uses this measure. The measure has been endorsed for five years and I am concerned that the only data source provided in this document is from the institution that is the developer. i agree with the staff rating on usability and use as LOW.

Criterion 5: Related and Competing Measures

Related or competing measures

- No related or competing measures identified.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0713
Measure Title: Ventriculoperitoneal (VP) shunt malfunction rate in children
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 5/24/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☒ Health outcome: 30-day malfunction rate for hospitals that perform cerebrospinal ventriculoperitoneal shunt operations in children between the ages of 0 and 18 years

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Ventricular shunt (VP) malfunction is related to the quality of the VP shunt placement.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Ventricular shunt malfunction places children at risk for potentially irreversible neurologic system deficits and death if not treated promptly. Shunt malfunction treatment is associated with the need for hospitalization and re-operation. The hospitalization itself is disruptive to the child and family, which may lead to impaired quality of life. The need for re-operation places the child at additional risk for central nervous system infection and other adverse events.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION N/A

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

N/A

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

N/A

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

N/A

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION N/A

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):
N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.
N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:
N/A
1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)
N/A
1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**
N/A
Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE N/A
1a.6.1. Citation (*including date*) and **URL** (*if available online*):
N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**
N/A
Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE
If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.
N/A
1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?
N/A
1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:
N/A
1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.
N/A
1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

N/A
QUANTITY AND QUALITY OF BODY OF EVIDENCE
1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)
N/A
1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)
N/A

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

N/A

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

N/A

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

N/A

_____

1a.8 OTHER SOURCE OF EVIDENCE

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

N/A

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

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

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

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

Evidence-635996758102740089.docx

1b. Performance Gap

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

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

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

Ventricular shunt malfunction places children at risk for potentially irreversible neurologic system deficits and death if not treated promptly. Shunt malfunction treatment is associated with the need for hospitalization and re-operation. The hospitalization itself is disruptive to the child and family, which may lead to impaired quality of life. The need for re-operation places the child at additional risk for central nervous system infection and other adverse events.

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

Shunt malfunction rates vary widely among different institutions.

During the most recent three-year reporting period (CY12 Q2 - CY15 Q1), Boston Children's Hospital had 46 eligible cases of ventricular shunt placement, of which 2 (4.35%) experienced a malfunction within 30 days. The SMR for these cases was 1.42 [95% CI (0.16, 5.11)]. This is not significantly different from the null value of 1.

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

N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Variation in shunt malfunction rates has not been assessed across population groups due to the fact that these data are not required to be collected at the institutions utilizing the measure.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Predicting shunt failure in children: should the global shunt revision rate be a quality measure? Rossi, N., Khan, N., Jones, T., Lepard, J., McAbee, J., Klimo, P. Jr. J Neurosurg Pediatr. 2015 Nov 6:1-11.

A multi-institutional, 5-year analysis of initial and multiple ventricular shunt revisions in children. Berry JG, Hall MA, Sharma V, Goumnerova L, Slonim AD, Shah SS. Neurosurgery. 2008 Feb;62(2):445-53; discussion 453-4.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Leading cause of morbidity/mortality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Children who require on-going cerebrospinal fluid diversion with a ventricular shunt have a major risk of morbidity and mortality. These children are experiencing high rates of life-threatening shunt malfunction.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Revision rate of pediatric ventriculoperiteneal shunts after 15 years. J Neurosurg Pediatr. 2013 Jan;11(1):15-9.

2. Infection rates following initial cerebrospinal fluid shunt placement across pediatric hospitals in the United States. Clinical article. Simon TD, Hall M, Riva-Cambrin J, Albert JE, Jeffries HE, Lafleur B, Dean JM, Kestle JR; Hydrocephalus Clinical Research Network. J Neurosurg Pediatr. 2009 Aug;4(2):156-65.

3. A multi-institutional, 5-year analysis of initial and multiple ventricular shunt revisions in children. Berry JG, Hall MA, Sharma V, Goumnerova L, Slonim AD, Shah SS. Neurosurgery. 2008 Feb;62(2):445-53; discussion 453-4.

4. A multicenter study of factors influencing cerebrospinal fluid shunt survival in infants and children. Shah SS, Hall M, Slonim AD, Hornig GW, Berry JG, Sharma V. Neurosurgery. 2008 May;62(5):1095-102; discussion 1102-3.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

evaluated against the remaining criteria.

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

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

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** ICD9_to_10_mapping_PHIS-VPShunt-635996755578611549.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The number of initial ventriculoperitoneal (VP) shunt placement procedures performed on children between the ages of 0 and 18 years of age that malfunction and result in shunt revision within 30 days of initial placement.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Within 30 days of initial VP shunt placement.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of cases of initial VP shunt placement (ICD-10 procedure codes 0016072, 0016073, 00160J2, 00160J3 00160K2, 00160K3, 0016372, 0016373, 00163J2, 00163J3, 00163K2, 00163K3, 0016074, 00160J4, 00160K4, 0016374, 00163J4, 00163K4, 0W110J9, 0W110JB, 0016076, 00160J6, 00160K6, 0016376, 00163J6, 00163K6, 0W110JG, 0W110JJ, 0016077, 00160J7, 00160K7, 0016377, 00163J7, 00163K7 (either as a primary of secondary procedure)) among patients between the ages of 0 and 18 years at the time of placement resulting in a malfunction characterized by a shunt revision within 30 days of initial procedure.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The total number of initial cerebrospinal VP shunt procedures performed on children between the ages of 0 and 18 years.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Children's Health

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
The total number of initial VP shunt placements (ICD-10 procedure codes 0016072, 0016073, 00160J2, 00160J3 00160K2, 00160K3, 0016372, 0016373, 00163J2, 00163J3, 00163K2, 00163K3, 0016074, 00160J4, 00160K4, 0016374, 00163J4, 00163K4, 0W110J9, 0W110JB, 0016076, 00160J6, 00160K6, 0016376, 00163J6, 00163K6, 0W110JG, 0W110JJ, 0016077, 00160J7, 00160K7, 0016377, 00163J7, 00163K7 (either as a primary of secondary procedure)) among patients between the ages of 0 and 18 years at the time of procedure. Patients also have no evidence of VP shunt placement or removal in the year prior to their initial procedure.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients with evidence of VP shunt placement or removal in the year prior to their index procedure are excluded.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Patients with evidence of VP shunt placement (ICD-10 procedure codes 0016072, 0016073, 00160J2, 00160J3 00160K2, 00160K3, 0016372, 0016373, 00163J2, 00163J3, 00163K2, 00163K3, 0016074, 00160J4, 00160K4, 0016374, 00163J4, 00163K4, 0W110J9, 0W110JB, 0016077, 00160J7, 00160K7, 0016377, 00163J7, 00163K7 (either as a primary of secondary procedure)) or malfunction (identified by ICD-10 procedure codes(either as a primary of secondary procedure) 00W60JZ, 00W63JZ, 00W64JZ (Revision of Synthetic Substitute in Cerebral Ventricle: Open Approach, Percutaneous Approach, Percutaneous Endoscopic Approach), or the combination of codes 00P60JZ, 00P63JZ, 00P64JZ (Removal of Synthetic Substitute from Cerebral Ventricle: Open Approach, Percutaneous Approach, Percutaneous Endoscopic Approach) and one of the following: 0016072, 0016073, 00160J2, 00160J3 00160K2, 00160K3, 0016372, 0016373, 00163J2, 00163J3, 00163K2, 00163K3, 0016074, 00160J4, 00160K4, 0016374, 00163J4, 00163K4, 0W110J9, 0W110JB, 0016076, 00160J6, 00160K6, 0016376, 00163J6, 00163K6, 0W110JG, 0W110JJ, 0016077, 00160J7, 00160K7, 0016377, 00163J7, 00163K7, 0016078, 00160J8, 00160K8, 00163J8, 00163K8, during the same admission in the year prior to their initial procedure are excluded.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
No Stratification is done with the data.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
We used logistic regression models to determine the risk adjustment variables.

The predicted value for each case is computed using a logistic regression model with covariates for with age at insertion (0-30 d, 31-365 d, and 1 y), congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity and spina bifida. The reference population used in the regression is the PHIS database from 2008-2010.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The measure is a 30-day VP shunt malfunction rate defined as the proportion of shunt revisions within 30 days over the number of initial cerebrospinal VP shunt placement procedures performed on children between the ages of 0 and 18 years. In order to stabilize the rates due to small number of events, the measure will be presented as a 3-year rolling rate. The benchmark for each year is the mean VP malfunction rate of all participating pediatric hospitals in the Pediatric Health Information System PHIS dataset.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not Specified. Because this is administrative data all eligible cases will be included. The rate will be presented as a 3 year rolling rate in order to account for fluctuations due to small number of events.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Pediatric Health Information System (PHIS):
PHIS is an administrative database that contains inpatient, emergency department and ambulatory surgery data from 42 not-for-profit, tertiary care pediatric hospitals in the United States. These hospitals are affiliated with the Child Health Corporation of America. Data quality and reliability are assured through a joint effort between the Child Health Corporation of America and participating hospitals.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

N/A
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** testing_attachment_-1--636005576464869055.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0713
Measure Title: Ventriculoperitoneal (VP) shunt malfunction rate in children
Date of Submission: 5/24/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects

the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

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

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

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

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

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

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

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☐ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 42T	☐ other: 42T

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

Data were obtained from BCH's Electronic Health Record (EHR) and administrative claims database. Benchmarking data were obtained from the Pediatric Health Information System (PHIS) database, which is compiled by the Child Health Corporation of America and contains clinical and financial information on patient admissions from 43 free-standing children's hospitals.

1.3. What are the dates of the data used in testing? CY12 Q2 – CY15 Q1

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice

☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 42T	☐ other: 42T

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)
Nine PHIS hospitals against which to benchmark were chosen based on the perceived similarity of their patient mix to BCH. They are: Birmingham, Philadelphia, Chicago, Dallas, Houston, LA, Palo Alto, St. Louis, and Seattle. The combined SMR of BCH and these hospitals serves as the benchmark.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
During the most recent three-year reporting period (CY12 Q2 - CY15 Q1), Boston Children's Hospital had 46 eligible cases of ventricular shunt placement, of which 2 (4.35%) experienced a malfunction within 30 days.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.
No differences in data used for different aspects of testing were found.

1.8. What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
N/A – we adjusted for clinical factors (age at insertion (0-30 d, 31-365 d, and =1 y), congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity and spina bifida)

2a2. RELIABILITY TESTING
***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*
Formal analysis of reliability/repeatability has not been performed due to the reliance on ICD-9/ICD-10 codes to calculate this measure.

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

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

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

N/A

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

2b2. VALIDITY TESTING
2b2.1. What level of validity testing was conducted? (*may be one or both levels*)
☒ **Critical data elements** (*data element validity must address ALL critical data elements*)
☐ **Performance measure score**
☐ **Empirical validity testing**
☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
Identification of ICD-10 codes for the Measure, "Ventriculoperitoneal (VP) Shunt Malfunction Rate in Children"

1. Experts involved in the ICD-9 to ICD-10 conversion:
   - Dr. Liliana Goumnerova, MD, FRCSC, Director, Pediatric Neurosurgical Oncology, Associate Professor, Boston Children's Hospital
   - Joseph Madsen, MD, Director of Epilepsy Program, Boston Children's Hospital
   - Sara Gernigan, MD, MPH, University of Miami Health System
2. 3M ICD-10 Code Translation Tool

Stakeholder comments:

Thanks for the chance to review your thoughtful work on our Shunt Outcomes Project. After reviewing the translation procedure from ICD 9 to ICD 10 for the VP shunt malfunction measure, it is my opinion that the original definitions of the population (numerator and denominator) have been maintained. To the best of my understanding, with this translation all appropriate and qualifying patients are still captured whether using ICD9 or ICD10 procedure codes (Dr. Madsen, MD).

The goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

In 2010, we tested the ICD-9 coding algorithm for shunt malfunction against shunt malfunction as determined by medical record review within two freestanding children's hospitals participating in the PHIS database, including Boston Children's Hospital (BCH) and another institution. All shunt procedures at BCH were reviewed (n=52). At the second institution, all of the five ICD9-coded shunt malfunctions were reviewed along with 55 randomly selected procedures with no malfunction codes (25% sample). The sensitivity, specificity, predicted positive value (PPV), and predicted negative value (PNV) were calculated overall and within each institution.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Among the 107 charts reviewed at the two institutions, the ICD-9 coding algorithm correctly identified 9 of the 12 malfunctions found by chart review resulting in a sensitivity of 0.75. The coding algorithm had high specificity (0.96) and a low false negative rate (3%). Performance of the coding algorithm was better at BCH than the second institution; however the differences did not reach statistical significance, perhaps due to small sample sizes.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Our results demonstrate that ICD-9-coded administrative databases, such as the PHIS dataset, can be used to evaluate VP shunt malfunction with acceptable sensitivity and high specificity. The high specificity of the algorithm provides the desirable property of guarding against over-estimation of the VP shunt malfunction rate.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 6 risk factors**

☐ **Stratification by 42T risk categories**

☐ **Other,** 42T

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

After both univariate and logistic regression analysis of the many variables that were suggestive of having a higher risk for having a shunt malfunction however no model stood apart. We tested different combinations (age at insertion (0-30 d, 31-365 d, and =1 y), congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity and spina bifida) of the variables that ended up in the final model. Previous published material has shown that these factors affect malfunction rates. So in the end we used a risk model with variables that were included age at insertion (0-30 d, 31-365 d, and =1 y), congenital anomalies, intraventricular hemorrhage, low birth weight, prematurity and spina bifida was used to adjust the incidence of malfunction.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

N/A

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
N/A

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A
Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:
The risk adjustment models were used to calculate the expected shunt malfunction rate for each hospital, adjusting for case mix.

Multivariate Analysis of Risk Factors for Shunt Malfunction within 30-days

Clinical Attribute	Model Coefficient ± Standard Error	Odds Ratio (95% CI)	p-value
Intercept	-2.65 ± 0.17	--	--
Age at insertion			
0-30 d	-- --	Reference	--
31-365 d	-0.47 ± 0.18	0.63 (0.44-0.89)	<0.001
≥1 y	-0.11 ± 0.20	0.90 (0.61-1.31)	0.56
Congenital Anomalies	-0.08 ± 0.15	0.92 (0.68-1.24)	0.58
Intraventricular hemorrhage	-0.20 ± 0.25	0.82 (0.50-1.34)	0.42
Low birth weight	0.48 ± 0.54	1.62 (0.56-4.67)	0..37
Prematurity	-0.12 ± 0.52	0.9 (0.32-2.49)	0.83
Spina Bifida	0.19 ± 0.20	1.2 (0.82-1.77)	0.34

Model c-statistic: 0.576

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
The results of the risk adjusted analysis showed no significant differences between the 6 identified risk factors in the odds of having shunt malfunction.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

PHIS Database 2012 through March 31 2015; 1,121 procedures performed at 10 selected institutions. Trend analysis is based on 3 year intervals. The combined VP malfunction rate of our institution (Children's Hospital Boston) and other PHIS participating hospitals serve as the benchmark. Meaningful differences between CHB and the benchmark will be assessed at 3 year intervals.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

During the most recent three-year reporting period (CY12 Q2 - CY15 Q1), Boston Children's Hospital had 46 eligible cases of ventricular shunt placement, of which 2 (4.35%) experienced a malfunction within 30 days. The SMR for these cases was 1.42 [95% CI (0.16, 5.11)]. This is not significantly different from the null value of 1.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The data from CY15Q1 indicate that our shunt malfunction rate is not significantly different from the null value of 1, given our patient case mix. These data are consistent with results from prior years, demonstrating the ability of the test to identify meaningful differences in performance across participating hospitals.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

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

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.e., *what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS
Missing data analysis has not yet done.

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.e., *what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
N/A

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Data generated as byproduct of care processes during care delivery (Data are generated and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition), Coding/abstraction performed by someone other than person obtaining original information (E.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or registry), Other
If other: Electronic medical record

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in a combination of electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

No modifications have been done to data collection. The data collection process relies on electronic medical record and is thus fairly straightforward.

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

N/A

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Public Reporting Pediatric National Surgical Quality Improvement Project (P-NSQIP) https://www.facs.org/quality-programs/pediatric Quality Improvement with Benchmarking (external benchmarking to multiple organizations) 10 participating institutions N/A Quality Improvement (Internal to the specific organization) Comprehensive Quality Report N/A

4a.1. **For each CURRENT use, checked above, provide:**

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

Pediatric National Surgical Quality Improvement Project (P-NSQIP). The success of the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP®) in providing hospitals with high-quality surgical outcomes data and methods to improve care has led to efforts to extend this robust quality improvement program to a larger number of hospitals and surgical specialties. The ACS has collaborated with the American Pediatric Surgical Association (APSA) to develop a pediatric version of ACS NSQIP. The program is open to all pediatric hospitals, including freestanding general acute care children's hospitals, children's hospitals within a larger hospital, specialty children's hospitals, or general acute care hospitals with a pediatric wing that want to collect reliable clinical data including 30-day outcomes. - See more at: https://www.facs.org/quality-programs/pediatric/program-specifics#sthash.4VV0lHVX.dpuf

QI with benchmarking. 10 pediatric hospitals, including BCH, which are members of the Pediatric Hospital Information System (PHIS). In CY12 Q2 - CY15 Q1, there were a combined cohort of 1,121 shunts and 37 shunt malfunction cases.

QI internal. The measure is used internally at BCH for quality improvement. It is included in the Comprehensive Quality Report, which summarizes the results of a hospital-wide initiative to measure quality of care. The results are presented bi-annually.

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

The shunt malfunction rate has remained close to the benchmark of 1 for the past 12 quarters (CY12 Q2 - CY 15 Q1). 3 Cases of shunt malfunction that were eligible occurred in 211 and 2012. We have not had any eligible cases since December 2012 through Q3 CY14.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
We continue with our current practice in an attempt to maintain our malfunction rate at the expected rate (value of 1), therefore we are not planning changes to our performance models. We are still working on addressing the issue of accurate stratification of patients and our expectation is that once we have better model development, we will see further changes.

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No major unintended consequences were identified during testing. Of note, PHIS is an administrative database and can be subject to coding inaccuracies and limitations. We are able to match our institution's cases in the PHIS database with our internal data system in order to assess accuracy.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Boston Children's Hospital, Center for Patient Safety and Quality Research
Co.2 **Point of Contact:** Maria, Jorina, Maria.Jorina@childrens.harvard.edu, 617-919-3613-
Co.3 **Measure Developer if different from Measure Steward:** Boston Children´s Hospital, Center for Patient Safety and Quality Research
Co.4 **Point of Contact:** Maria, Jorina, Maria.Jorina@childrens.harvard.edu, 617-919-3613-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Liliana Goumnerova, MD
Sara Jernigan, MD, MPH
Dionne Graham, PhD
All three members were involved in the original measure development and testing.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2006
Ad.3 **Month and Year of most recent revision:** 09, 2014
Ad.4 **What is your frequency for review/update of this measure?** Bi-annual review
Ad.5 **When is the next scheduled review/update for this measure?** 09, 2016

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

VP Shunt Malfunction Risk Factors – Literature review

1. **Age at insertion**
    a. Simon T, Whitlock K, Riva-Cambrin J, et al. Revision surgeries are associated with significant risk of subsequent cerebrospinal fluid shunt infection. *Pediatr Infect Dis J*. 2012. 31:551-556.
2. **Prematurity**
    a. McGirt M, Zaas A, Fuchs H, et al. Risk Factors for Pediatric Ventriculoperitoneal Shunt Infection and Predictors of Infectious Pathogens. *Clin Infect Dis*. 2003. 36(7):858-862.
3. **Intraventrocular hemorrhage, low birth weight**
    a. Jernigan S, Berry G, Graham D, Goumnerova L. The comparative effectiveness of ventricular shunt placement versus endoscopic third ventriculostomy for initial treatment of hydrocephalus in infants. *J Neurosurg Pediatrics*. 2014. 13:295-300.
4. **Spina bifida**
    a. Berry J, Hall M, Sharma V, et al. A multi-institutional, 5-year analysis of initial and multiple ventricular shunt revisions in children. *Neurosurgery*. 2008. 62(2):445-453.
    b. Caldarelli M, Rocco C, La Marca F. Shunt complications in the first postoperative year in children with meningomyelocele. *Child's Nerv Syst*. 1996. 12(12):748-754.
5. **Congenital anomalies**
    a. Kebriaei M., Shoja M., Salinas S, et al. Shunt infection in the first year of life. *J Neurosurg Pediatr*. 2013. 12:44-48.
    b. Riva-Cambrin J., Kestle J., Holubkov R, et al. Risk factors for shunt malfunction in pediatric hydrocephalus: a multicenter prospective cohort study. *J Neurosurg Pediatr*. 2016. 17:382-390.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1519
De.2. **Measure Title:** Statin Therapy at Discharge after Lower Extremity Bypass (LEB)
Co.1.1. **Measure Steward:** Society for Vascular Surgery
De.3. **Brief Description of Measure:** Percentage of patients aged 18 years and older undergoing infrainguinal lower extremity bypass who are prescribed a statin medication at discharge. This measure is proposed for both hospitals and individual providers.
1b.1. **Developer Rationale:** Based on the data summarized in this application, this quality measure will be associated with decreased perioperative morbidity and mortality from major adverse cardiac events including stroke, myocardial infarction, and death. The data also suggest a potential association between perioperative statin use and improved bypass graft patency.

Patients who require LEB have advanced peripheral arterial disease and meet guidelines for secondary prevention with statins. Many of these patients have not received adequate management of PAD risk factors. The episode of care associated with LEB provides an opportunity to initiate statin therapy in these patients in order to improve survival and reduce cardiovascular complications following the procedure.

S.4. **Numerator Statement:** Patients undergoing infrainguinal lower extremity bypass who are prescribed a statin medication at discharge.
S.7. **Denominator Statement:** All patients aged 18 years and older undergoing lower extremity bypass as defined above who are discharged alive, excluding those patients who are intolerant to statins.
S.10. **Denominator Exclusions:** Chart documentation that patient was not an eligible candidate for statin therapy due to known drug intolerance, or patient died before discharge.

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Facility

Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary or **Summary of prior review in 2010**

Clinical guidelines cited in support of the measure include:

- TransAtlantic Inter-Society Consensus guideline, "In symptomatic PAD patients, statins should be the primary agents to lower LDL cholesterol levels to reduce the risk of cardiovascular events. Grade: A (Based on the criterion of at least one randomized, controlled clinical trial as part of the body of literature of overall good quality and consistency addressing the specific recommendation – cited as AHRQ guidance)
- ACC/AHA guidelines, "Treatment with a hydroxymethyl glutaryl (HMG)coenzyme-A reductase inhibitor (statin) medication is indicated for all patients with PAD to achieve a target LDL cholesterol level of less than 100 mg per dL.(Class I, Level of Evidence: B) and "Treatment with an HMG coenzyme-A reductase inhibitor (statin) medication to achieve a target LDL cholesterol level of less than 70 mg per dL is reasonable for patients with lower extremity PAD at very high risk of ischemic events." (Class IIa: Level of Evidence: B)
- The developer cites several studies that support the relationship between statin prescription and decreased morbidity and mortality.
- Additional sources of evidence are listed here.

Changes to evidence from last review

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

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

Updates:

The developer submitted updated evidence for this measure:

- T.R. Vogel, V.Y. Dombrovskiy, E.L. Galiñanes. **Preoperative Statins and Limb Salvage After Lower Extremity Revascularization in the Medicare Population.** Journal of Vascular Surgery, Vol. 59, Issue 3, p873. Published in issue: March 2014
- Olaf Schouten, Sanne E. Hoeks, Michiel T. Voute, Eric Boersma, Hence J. Verhagen, Don Poldermans. **Long-term Benefit of Perioperative Statin Use in Patients Undergoing Vascular Surgery: Results from the DECREASE III Trial.** Journal of Vascular Surgery, Vol. 53, Issue 6, p20S–21S. Published in issue: June 2011

Questions for the Committee:

- *Does the evidence provided substantiate the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm

Systematic review (Box 3) → Quantity, Quality, and Consistency of evidence (Box 4) → Moderate certainty that the net benefit is strong (Box 5a) → High

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

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

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

improvement.

- In the previous consideration of this measure, the Committee noted that performance rates improved from 41% to 79%. Rates were still short of the 90% goal set by SVS and VSGNE.
- Data submitted for this submission show a level of improvement in 2010, then a slight decline, before improving to 79.1% in 2015:

Rate By Year

2010	2011	2012	2013	2014	2015
81.3%	78.5%	77.7%	77.8%	77.6%	79.1%

- In 2016, an analysis of VQI registry for the time period 2010-2015 found that of 23,410 infrainguinal LEB procedures across 188 centers and 914 providers, the median rate of statins prescribed at discharge was 77.8% with an interquartile range of 69.7% to 84.5%. The developer reports that while the data have shown improvement, their continues to be a performance gap.

Disparities

- In a 2016 analysis of the VQI registry for the time period 2010-2015, the developer reports that of 23,410 infrainguinal LEB procedures(across 188 centers and 914 providers), patients younger than 60 and older than 80 were less likely to be on statins than those aged 60-79 years. The developer states that there was a slight difference (2%) by gender and by Hispanic versus non-Hispanic.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is there expected variation in performance if reported at the physician level versus at the facility level?*
- *Does the Committee know of any additional disparities in statin prescription?*

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

Committee pre-evaluation comments

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

1a.

- Process measure but strong validity
- Wide gap in performance
- Should be measurable in multiple datasets/registries
- Easily enabled e-measure
- This process measure is up for Maintenance of Endorsement. Originally approved Jan 31, 2012. There is no new evidence; the new evidence provided by the developer are abstracts that discuss articles that were published prior to the 2012 approval. That said, there is no change in the evidence; consensus guidelines call for all symptomatic PAD patients be started on a statin and ACC/AHA guideline recommend all PAD patients be treated with a statin. Grades A and B respectively.
- There is a direct relationship between perioperative statin use and subsequent mortality as documented in several referenced RCTs. One of the newly added studies suggests that pre-operative initiation is superior to postoperative in terms of long term survival.

1b.

- Initial performance improvement of 41% to 79%. Improvement occurred in 1st year of measure and improvement has been flat since. 2016 analysis of VQI registry (23,410 patients 2010-2015) indicate a 77.8% compliance at discharge with interquartile range of 69.7-84.5%.
- There is still room for improvement when considering exclusion criteria and performance of the measure.
- Initial performance improved from 41% to 79%, but performance has plateaued in the last 5 years. The goal set by SVS was set at 90% use. Older and younger patients are less likely to get statins, as are Hispanic vs non-Hispanic (slightly)

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data sources: Electronic Clinical Data: Registry. Registries used for this measure are The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE). VQI and VSGNE include hospitalization details and symptom status.

Specifications:

- This measure is specified at the Clinician: Group/Practice and Clinician: Individual, Facility level for use in hospital/acute care facilities.
- The numerator includes patients undergoing infrainguinal lower extremity bypass (LEB) who are prescribed a statin at discharge.
- The denominator includes all patients aged 18 and older undergoing LEB who are discharged alive. Included CPT codes are found here.
- Exclusions include patients with known drug intolerance and patients who died before discharge. These data are captured in the SVS VQI and VSGNE registries.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- No new testing data was provided since the last endorsement. Testing presented during the previous review of this measure is presented below.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- A random sample of 100 patient records were reviewed, representing 5 LEB procedures from 5 different hospitals based on data collected during the "past 2 years". (Note that the 'past 2 years' refers to the time period prior to when the measure was first endorsed.)
- Hospital reporting is proposed for every 12 months based on sufficient volume.
- Annual reporting of the last 50 consecutive procedures for surgeons (which may span more than one year) with suppression of <10 procedures is recommended.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007.
- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched

to patient data within the registry to compare discharge status (alive vs dead).
- Any discrepancies were further evaluated based on a medical record audit.
- Kappa statistic was used to judge reliability of the data.

Results of reliability testing

- Data element validity testing was used to support the reliability of the measure.
- The Kappa statistic indicated strong agreement for identification of the correct procedure (LEB) performed (1.0), statin prescribed at discharge (0.80), hospital mortality (.91), age for 18 and older (100% agreement, 1.0) and intolerant to statins (1.0).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*

Guidance from the Reliability Algorithm:

Precise specifications (Box 1) → Empirical reliability testing (Box 2) →Patient level data validity (Box 3) → (Box 10 of validity algorithm) →Appropriate method to assess data elements (Box 11) → Moderate certain that data used in the measure are valid (Box 12a) →Highest possible rating is moderate.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- No new testing data was provided since the last endorsement. A summary of testing is presented below.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- A random sample of 100 patients records were examined representing 5 LEB procedures from 5 hospitals based on data collected from the "past 2 years". (Note that the 'past 2 years' refers to the time period prior to the first endorsement.)
- Validity testing of statin prescribed at discharge used the medical record as the gold standard.
- Chart abstraction was completed with comparison to registry data. Medication lists from both the discharge

summary and discharge orders were compared to confirm validity. Operation type using the operative report was compared with the progress note in the medical record.
- The developer states that patient age and hospital mortality have face validity.

Validity testing results:
- The developer reports 100% agreement was found between statin prescribed at discharge on the discharge summary and on the discharge orders; and between the procedure type documented in the operative note and in the progress notes.

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
- Patients are excluded if the patient has known drug intolerance, or the patient died before discharge.
- Exclusions analysis was completed for 2,496 patients in the VSGNE registry who had undergone infrainguinal LEB between 2003-2010.
- Analysis showed that 2% of exclusions were those patients who died in the hospitals, and another 2% for those who were alive but intolerant to statins.
- Of the remaining patients, 73% were discharged on statins.
- Numbers for the subsets below not provided.
  - Across 13 hospitals, the median statin prescription rate at discharge was 73% (interquartile range of 69% to 80%).
  - Across 63 individual providers, median statin prescription rate at discharge was 75% (interquartile range of 66% to 84%).

Questions for the Committee:
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Exclusion analysis documented above, related to 2,496 patients in the VSGNE registry who had LEB between 2003-2010, showed that 2% of exclusions were those patients who died in the hospitals, and another 2% for those who were alive but intolerant to statins.
- Of the remaining patients, 73% were discharged on statins.
  - Across 13 hospitals, the median statin prescription rate at discharge was 73% (interquartile range of 69% to 80%).
  - Across 63 individual providers, median statin prescription rate at discharge was 75% (interquartile range of 66% to 84%).
- The developer reports using standard statistical analysis to determine 95% confidence interval for hospital and providers to determine practical difference from the mean.

Question for the Committee:
- *Does this information provided identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- The developer states that other data sources are not available for testing.

2b7. Missing Data

- The developers report less than 2% of data were missing from both the VQI and VSGNE registries.

Guidance from the Validity Algorithm
Precise specifications (Box 1) → Potential threats to validity mostly assessed (Box 2) → Empirical validity testing conducted (Box 3) →Testing at the level of the measure score not conducted (Box 6) → Data element testing conducted (Box 10) →Appropriate assessment of data elements (Box 11) → certainty (Box 12a)

Highest possible rating is Moderate.

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

- The data elements are described clearly, but do require participation in a registry for implementation. Generic specifications (such as could be used using administrative claims alone) are not provided. Lack of these data would inhibit consistent implementation outside of a registry.

2a.2

- Kappa statistic indicates strong agreement
- reliability testing was done at the previous endorsement. I note that the key numerator metric showed 80% agreement (Kappa stat) which is acceptable in research studies but may be sub par for quality measures tied to reimbursement.

2b.1

- A small sample of 100 patients from 5 institutions collected over 2 years was reported. these were 100% compared to gold standard of medical record testing. Sample size appears small considering 23K+ patients were evaluated from 2010-2015
- They are consistent with the evidence

2b.2

- The validity testing seems to involve testing various elements of the medical record against each other (eg do the discharge orders medication lists match the discharge summary's). I'm not sure this is 'validity'. it seems to me that the metrics reported in 'reliability' are more appropriate for 'validy' testing. The measure has face validity.

2b.3-7.

- < 2% missing
- There are a small number of exclusions that comprise about 4% of cases. No risk adjustment is needed. There appear to be meaningful differences in quality with little missing data.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data are coded by someone other than the person obtaining the original information.
- Data are pulled from VSGNE. The developer reports VSGNE has been tracking statin usage since 2003 and have not experienced any difficulty obtaining statin usage data. There have been no problems with collecting data from VQI.
- Developer reports missing perioperative statin prescription has been less than 2% for both VSGNE and VQI.
- There are no fees or licenses needed to use the measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are there fees to belong to the registry?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

- VSGNE and Vascular Quality Initiative (VQI) report <2% missing data
- Feasibility is good within registry participants. Apparently the use of the registry is growing. It doesn't seem to be feasible outside of the related registries.

Criterion 4: Usability and Use

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

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

Current uses of the measure

- Centers for Medicaid and Medicare Services, Physicians Quality Reporting System (Measure 257)

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- PQRS is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). PQRS measures are used for public reporting on the Physician Compare website and for the quality component of the Value-Based Payment Modifier (VBPM).

Improvement results

- The measure saw some improvement from 2010. Improvement declined after being included in PQRS and the VQI. In 2015 improvements were again noted.

Rate By Year

2010	2011	2012	2013	2014	2015
81.3%	78.5%	77.7%	77.8%	77.6%	79.1%

- The developer notes that the number of procedures captured in the VQI registry has increased in each of these years.

Unexpected findings (positive or negative) during implementation

- The developer acknowledges that it is possible to miss or inaccurately code statins and has overcome this by providing a list of generic and trade names of known statin medications.

Potential harms

- The developer did not report whether there were potential harms.

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Reported through the Physician Compare website; paytments and payment adjustments are made as report by PQRS. Data not provided
- it is reported as part of the PQRS system. No significant unintended consequences.

Criterion 5: Related and Competing Measures

Related or competing measures

- #0118 Anti-Lipid Treatment Discharge
- #0439 Discharged on Statin Medications
  - During the previous evaluation of this measure, the Committee stated that measures 0118 and 1519 were related in terms of therapy used; however, they involve different procedures and different patient populations and are reasonably aligned thus no further action was recommended.
  - #0439 has been recommended for endorsement with reserve status by the Neurology Standing Committee.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1519 NQF Project: Surgery Endorsement Maintenance 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

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

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:
As summarized above, this quality measure will be associated with decreased perioperative morbidity and mortality from major adverse cardiac events including stroke, myocardial infarction, and death in patients undergoing lower extremity bypass. The data also suggest a potential association between perioperative statin use and improved bypass graft patency.

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
Please see the summary of the data presented in 1.a.3.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*:

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Data obtained from randomized prospective controlled trials.
1. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7-22.
2. Randomized trial of the effects of cholesterol-lowering with simvastatin on peripheral vascular and other major vascular outcomes in 20,536 people with peripheral arterial disease and other high-risk conditions. J Vasc Surg 2007;45:645-54
3. Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. N Engl J Med 2009;361:980-9.

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Level 1.

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
1. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7-22.
2. Randomized trial of the effects of cholesterol-lowering with simvastatin on peripheral vascular and other major vascular outcomes in 20,536 people with peripheral arterial disease and other high-risk conditions. J Vasc Surg 2007;45:645-54; discussion 53-4.
3. Schanzer A, Hevelone N, Owens CD, Beckman JA, Belkin M, Conte MS. Statins are independently associated with reduced mortality in patients undergoing infrainguinal bypass graft surgery for critical limb ischemia. J Vasc Surg 2008;47:774-81.

4. Feringa HH, Karagiannis SE, van Waning VH, Boersma E, Schouten O, Bax JJ, et al. The effect of intensified lipid-lowering therapy on long-term prognosis in patients with peripheral arterial disease. J Vasc Surg 2007;45:936-43.
5. Ward RP, Leeper NJ, Kirkpatrick JN, Lang RM, Sorrentino MJ, Williams KA. The effect of preoperative statin therapy on cardiovascular outcomes in patients undergoing infrainguinal vascular surgery. Int J Cardiol 2005;104:264-8.
6. Kertai MD, Boersma E, Westerhout CM, van Domburg R, Klein J, Bax JJ, et al. Association between long-term statin use and mortality after successful abdominal aortic aneurysm surgery. Am J Med 2004;116:96-103.
7. Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. N Engl J Med 2009;361:980-9.
8. Poldermans D, Bax JJ, Kertai MD, Krenning B, Westerhout CM, Schinkel AF, et al. Statins are associated with a reduced incidence of perioperative mortality in patients undergoing major noncardiac vascular surgery. Circulation 2003;107:1848-51.
9. O´Neil-Callahan K, Katsimaglis G, Tepper MR, Ryan J, Mosby C, Ioannidis JP, et al. Statins decrease perioperative cardiac complications in patients undergoing noncardiac vascular surgery: the Statins for Risk Reduction in Surgery (StaRRS) study. J Am Coll Cardiol 2005;45:336-42.
10. Christenson J. Preoperative lipid control with simvastatin reduces the risk for graft failure already 1 year after myocardial revascularization. Cardiovasc Surg 2001;9:33-43.
11. Abbruzzese TA, Havens J, Belkin M, Donaldson MC, Whittemore AD, Liao JK, et al. Statin therapy is associated with improved patency of autogenous infrainguinal bypass grafts. J Vasc Surg 2004;39:1178-85.
12. Henke PK, Blackburn S, Proctor MC, Stevens J, Mukherjee D, Rajagopalin S, et al. Patients undergoing infrainguinal bypass to treat atherosclerotic vascular disease are underprescribed cardioprotective medications: effect on graft patency, limb salvage, and mortality. Journal of Vascular Surgery 2004;39:357-65.

New Citations for 2016 Maintenance:

13. T.R. Vogel, V.Y. Dombrovskiy, E.L. Galiñanes. **Preoperative Statins and Limb Salvage After Lower Extremity Revascularization in the Medicare Population.** Journal of Vascular Surgery, Vol. 59, Issue 3, p873. Published in issue: March 2014

14. Olaf Schouten, Sanne E. Hoeks, Michiel T. Voute, Eric Boersma, Hence J. Verhagen, Don Poldermans. **Long-term Benefit of Perioperative Statin Use in Patients Undergoing Vascular Surgery: Results from the DECREASE III Trial.** Journal of Vascular Surgery, Vol. 53, Issue 6, p20S–21S. Published in issue: June 2011

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
Recommendation #2, Section B1.2.3 (Dormandy et al.)
"In symptomatic PAD patients, statins should be the primary agents to lower LDL cholesterol levels to reduce the risk of cardiovascular events (1)."

Section 2.6.1.1. (Hirsch et al)
"Treatment with a hydroxymethyl glutaryl (HMG)coenzyme-A reductase inhibitor (statin) medication is indicated for all patients with PAD to achieve a target
LDL cholesterol level of less than 100 mg per dL.(Level of Evidence: B)
1. Treatment with an HMG coenzyme-A reductase inhibitor (statin) medication to achieve a target LDL cholesterol level of less than 70 mg per dL is reasonable
for patients with lower extremity PAD at very high risk of ischemic events. (Level of Evidence: B"

1c.17 Clinical Practice Guideline Citation: 1. Dormandy JA, Rutherford RB. Management of peripheral arterial disease (PAD). TASC Working Group. TransAtlantic Inter-Society Consensus (TASC). J Vasc Surg 2000;31:S1-S296.
2. Hirsch AT, Haskal ZJ, Hertzer NR, Bakal CW, Creager MA, Halperin JL, et al. ACC/AHA 2005 Practice Guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease): endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart, Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation. Circulation 2006;113:e463-654.

1c.18 National Guideline Clearinghouse or other URL: NA

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: NA

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: NA

1c.24 Rationale for Using this Guideline Over Others: This quality measure will be associated with decreased perioperative morbidity and mortality from major adverse cardiac events including stroke, myocardial infarction, and death, in patients undergoing lower extremity bypass.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: 1c.26 Quality: 1c.27 Consistency:

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1519_Evidence_MSF5.0_Data_2016.doc

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Based on the data summarized in this application, this quality measure will be associated with decreased perioperative morbidity and mortality from major adverse cardiac events including stroke, myocardial infarction, and death. The data also suggest a potential association between perioperative statin use and improved bypass graft patency.

Patients who require LEB have advanced peripheral arterial disease and meet guidelines for secondary prevention with statins. Many of these patients have not received adequate managment of PAD risk factors. The episode of care assoicated with LEB provides an opportunity to initiate statin therapy in these patients in order to improve survival and reduce cardiovascular complications following the procedure.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Current guidelines support the use of statin therapy in all PAD patients with a target LDL level of less than 100 mg/dL (<70 mg/dL for patients deemed at very high risk).18 Because of the pleiotrophic effects of statins, PAD guidelines recommend that all PAD patients be treated, independent of LDL level.

However, a significant percentage of patients undergoing lower extremity bypass are not on statin therapy before or after surgery. In the PREVENT III trial referenced above, only 46% of patients were on statin therapy prior to surgery and only 45% of patients were prescribed statin therapy on hospital discharge.8 In the Vascular Study Group of New England (VSGNE), a multicenter quality improvement consortium, data has been collected on 3,693 patients who have undergone LEB. Unpublished analyses of these data demonstrate that only 41% of patients were taking statins preoperatively before LEB in 2004. Through quality improvement efforts, this percentage of patients dischared on statins has increased to 79% during the first 6 months of 2010. However, this rate of statin use falls significantly short of the 90% goal set forth by this quality improvement group in 2008. This under-treatment of patients with PAD has been echoed by several other reports in the literature and provides substantial opportunity for improvement.19-21

Patients undergoing infrainguinal LEB in VSGNE were analyzed for this measure submission. There are 2496 patients in the registry who underwent infrainguinal LEB between 2003-2010. Of these, 2% died in hospital. Of those discharged alive, only 2% were intolerant to statins. Across 13 hospitals, the median statin prescribed at discharge rate was 73%, with an interquartile range of 69% to 80%. Across 63 individual providers, the median statin prescribed at discharge rate was 75%, with an interquartile range of 66% to 84%. SVS and VSGNE have set quality targets at 90%. These data demonstrate both significant variation and a significant performance gap.

In spring 2016, an analysis was conducted by the Vascular Quality Initiative (VQI) registry specifically to provide performance scores on this measure for maintenance. The analysis was conducted for the time period of 2010 - 2015. There were 23,410 infrainguinal LEB procedures included in the registry between 2010 - 2015. Across 188 centers with 914 providers, the median statin prescribed at discharge was 77.8% with an interquartile range of 69.7% to 84.5%. While the data demonstrate some improvement over the last five years, their continues to be significant variation and a significant performance gap.

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

1. Dormandy JA, Rutherford RB. Management of peripheral arterial disease (PAD). TASC Working Group. TransAtlantic Inter-Society Consensus (TASC). J Vasc Surg 2000;31:S1-S296.
2. Criqui MH, Langer RD, Fronek A, Feigelson HS, Klauber MR, McCann TJ, et al. Mortality over a period of 10 years in patients with peripheral arterial disease. N Engl J Med 1992;326:381-6.
3. McKenna M, Wolfson S, Kuller L. The ratio of ankle and arm arterial pressure as an independent predictor of mortality. Atherosclerosis 1991;87:119-28.
4. Howell MA, Colgan MP, Seeger RW, Ramsey DE, Sumner DS. Relationship of severity of lower limb peripheral vascular disease to mortality and morbidity: a six-year follow-up study. J Vasc Surg 1989;9:691-6; discussion 6-7.
5. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7-22.
6. Randomized trial of the effects of cholesterol-lowering with simvastatin on peripheral vascular and other major vascular outcomes in 20,536 people with peripheral arterial disease and other high-risk conditions. J Vasc Surg 2007;45:645-54; discussion 53-4.
7. Conte MS, Bandyk DF, Clowes AW, Moneta GL, Seely L, Lorenz TJ, et al. Results of PREVENT III: a multicenter, randomized trial of edifoligide for the prevention of vein graft failure in lower extremity bypass surgery. J Vasc Surg 2006;43:742-51; discussion 51.
8. Schanzer A, Hevelone N, Owens CD, Beckman JA, Belkin M, Conte MS. Statins are independently associated with reduced mortality in patients undergoing infrainguinal bypass graft surgery for critical limb ischemia. J Vasc Surg 2008;47:774-81.
9. Feringa HH, Karagiannis SE, van Waning VH, Boersma E, Schouten O, Bax JJ, et al. The effect of intensified lipid-lowering therapy on long-term prognosis in patients with peripheral arterial disease. J Vasc Surg 2007;45:936-43.
10. Ward RP, Leeper NJ, Kirkpatrick JN, Lang RM, Sorrentino MJ, Williams KA. The effect of preoperative statin therapy on cardiovascular outcomes in patients undergoing infrainguinal vascular surgery. Int J Cardiol 2005;104:264-8.
11. Kertai MD, Boersma E, Westerhout CM, van Domburg R, Klein J, Bax JJ, et al. Association between long-term statin use and mortality after successful abdominal aortic aneurysm surgery. Am J Med 2004;116:96-103.
12. Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. N Engl J Med 2009;361:980-9.
13. Poldermans D, Bax JJ, Kertai MD, Krenning B, Westerhout CM, Schinkel AF, et al. Statins are associated with a reduced incidence of perioperative mortality in patients undergoing major noncardiac vascular surgery. Circulation 2003;107:1848-51.
14. O´Neil-Callahan K, Katsimaglis G, Tepper MR, Ryan J, Mosby C, Ioannidis JP, et al. Statins decrease perioperative cardiac complications in patients undergoing noncardiac vascular surgery: the Statins for Risk Reduction in Surgery (StaRRS) study. J Am Coll Cardiol 2005;45:336-42.
15. Christenson J. Preoperative lipid control with simvastatin reduces the risk for graft failure already 1 year after myocardial

revascularization. Cardiovasc Surg 2001;9:33-43.
16. Abbruzzese TA, Havens J, Belkin M, Donaldson MC, Whittemore AD, Liao JK, et al. Statin therapy is associated with improved patency of autogenous infrainguinal bypass grafts. J Vasc Surg 2004;39:1178-85.
17. Henke PK, Blackburn S, Proctor MC, Stevens J, Mukherjee D, Rajagopalin S, et al. Patients undergoing infrainguinal bypass to treat atherosclerotic vascular disease are underprescribed cardioprotective medications: effect on graft patency, limb salvage, and mortality. Journal of Vascular Surgery 2004;39:357-65.
18. Hirsch AT, Haskal ZJ, Hertzer NR, Bakal CW, Creager MA, Halperin JL, et al. ACC/AHA 2005 Practice Guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease): endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart, Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation. Circulation 2006;113:e463-654.
19. Hirsch AT, Criqui MH, Treat-Jacobson D, Regensteiner JG, Creager MA, Olin JW, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. Jama 2001;286:1317-24.
20. McDermott MM, Mehta S, Ahn H, Greenland P. Atherosclerotic Risk Factors Are Less Intensively Treated in Patients with Peripheral Arterial Disease Than in Patients with Coronary Artery Disease. J Gen Intern Med 1997;12:209-15.
21. Mukherjee D, Lingam P, Chetcuti S, Grossman PM, Moscucci M, Luciano AE, et al. Missed opportunities to treat atherosclerosis in patients undergoing peripheral vascular interventions: insights from the University of Michigan Peripheral Vascular Disease Quality Improvement Initiative (PVD-QI2). Circulation 2002;106:1909-12.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
There are not published data regarding disparities in statin usage after infrainguinal bypass in different population groups. Such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region.

In spring 2016, an analysis was conducted by the Vascular Quality Initiative (VQI) registry specifically to provide performance scores on this measure for maintenance. The analysis was conducted for the time period of 2010 - 2015. There were 23,410 infrainguinal LEB procedures included in the registry between 2010 - 2015, from 188 centers with 914 providers reporting. The data did show that for those individuals whose age was either less than 60 years old or greater than 80 years old, the use of statins was slightly less (5.5% and 7%) versus those age 60 - 79 years old. The data also demonstrated a slight, 2% difference, by gender and a slight, 2% difference, by Hispanic versus non-Hispanic.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
None found

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, Frequently performed procedure, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
Patients who present with lower extremity ischemia bear a large systemic burden of atherosclerotic disease, and therefore face not only the immediate risk of limb loss1 but also an increased risk for cardiovascular events.2-4 The benefits of statin therapy for

cardiovascular risk reduction in the PAD population have been demonstrated in several studies, most notably the Heart Protection Study.5, 6 The Heart Protection Study (HPS) is the largest trial to assess the effects of statins on major morbidity and mortality. The investigators enrolled over 20,000 patients deemed to be at high risk for cardiovascular events and randomized them to receive either 40mg of simvastatin or placebo. On survival analysis, they demonstrated that treatment with a statin was significantly associated with a decrease in all-cause mortality (12.9% vs. 14.7%, p=.0003) and that this effect was primarily driven by the reduction in death from vascular causes (7.6% vs. 9.1%, p<.0001). A recently published subgroup analysis6 focusing specifically on patients with documented PAD (n=6748) did not include mortality data. However, the authors demonstrated a significant reduction in the rate of first major vascular event in the simvastatin treatment arm (relative reduction of 22%; p<.0001), when compared to placebo.

The PREVENT III trial was a prospective, randomized, double-blinded, multicenter trial designed to examine the efficacy of a novel pharmacologic agent (edifoligide) in preventing autogenous vein graft failure in 1404 patients who underwent infrainguinal vein bypass at 83 hospitals exclusively for the treatment of critical limb ischemia.7 This LEB trial, with its high-risk critical limb ischemia (CLI) population, provides another relevant database for examination of the role of statins. The salient finding from this study is that the use of statin drugs was associated with a significant one-year survival benefit in patients undergoing surgical bypass for CLI.8 The Kaplan-Meier analysis also suggested that the benefit continues to increase with time, and might be even greater with longer term follow-up. In these 1404 patients, those not receiving statins experienced a 40% increase in the risk of death at one year. This effect was demonstrated both in the propensity score weighted analysis (HR 1.40, CI 1.02-1.92), and in the Cox proportional hazards model (HR 1.47, CI 1.11-1.96). These findings are consistent with prior observational studies that have examined the effects of statins, albeit, in heterogeneous PAD populations.9-11 The largest of these observational studies, conducted by Feringa and colleagues, enrolled 1374 patients with PAD and followed them for a mean duration of 6.4 years. The authors demonstrated a strong independent association between statin use and all-cause mortality (HR 1.41 for non-users, p<0.0001).9

The DECREASE study randomized 497 patients who had not previously been treated with a statin to receive either 80 mg of extended-release fluvastatin or placebo once daily before undergoing major non-cardiac vascular surgery.12 On evaluation of the primary endpoint, statin therapy conferred a 45% decreased hazard ratio (10.8% versus 19%, p=0.01) for perioperative myocardial infarction. Furthermore, death from cardiovascular causes or myocardial infarction occurred in 4.8% of patients in the fluvastatin group and 10.1% of patients in the placebo group (hazard ratio, 0.47; 95% CI, 0.24 to 0.94; p= 0.03). Fluvastatin therapy was not associated with a significant increase in the rate of adverse events. Several additional studies in patients undergoing LEB have shown similar reductions in perioperative morbidity and mortality associated with statin use.10, 13, 14

Recent studies have also demonstrated a specific benefit in graft patency after LEB in patients on statin therapy.15-17 Abbruzzese et al observed that statin use was associated with improved secondary patency (3-fold increased risk compared to non-users) among 197 patients who had undergone lower extremity bypass using saphenous vein, in a single-center retrospective analysis.16

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Dormandy JA, Rutherford RB. Management of peripheral arterial disease (PAD). TASC Working Group. TransAtlantic Inter-Society Consensus (TASC). J Vasc Surg 2000;31:S1-S296.
2. Criqui MH, Langer RD, Fronek A, Feigelson HS, Klauber MR, McCann TJ, et al. Mortality over a period of 10 years in patients with peripheral arterial disease. N Engl J Med 1992;326:381-6.
3. McKenna M, Wolfson S, Kuller L. The ratio of ankle and arm arterial pressure as an independent predictor of mortality. Atherosclerosis 1991;87:119-28.
4. Howell MA, Colgan MP, Seeger RW, Ramsey DE, Sumner DS. Relationship of severity of lower limb peripheral vascular disease to mortality and morbidity: a six-year follow-up study. J Vasc Surg 1989;9:691-6; discussion 6-7.
5. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002;360:7-22.
6. Randomized trial of the effects of cholesterol-lowering with simvastatin on peripheral vascular and other major vascular outcomes in 20,536 people with peripheral arterial disease and other high-risk conditions. J Vasc Surg 2007;45:645-54; discussion 53-4.
7. Conte MS, Bandyk DF, Clowes AW, Moneta GL, Seely L, Lorenz TJ, et al. Results of PREVENT III: a multicenter, randomized trial of edifoligide for the prevention of vein graft failure in lower extremity bypass surgery. J Vasc Surg 2006;43:742-51; discussion 51.
8. Schanzer A, Hevelone N, Owens CD, Beckman JA, Belkin M, Conte MS. Statins are independently associated with reduced mortality in patients undergoing infrainguinal bypass graft surgery for critical limb ischemia. J Vasc Surg 2008;47:774-81.
9. Feringa HH, Karagiannis SE, van Waning VH, Boersma E, Schouten O, Bax JJ, et al. The effect of intensified lipid-lowering therapy on long-term prognosis in patients with peripheral arterial disease. J Vasc Surg 2007;45:936-43.
10. Ward RP, Leeper NJ, Kirkpatrick JN, Lang RM, Sorrentino MJ, Williams KA. The effect of preoperative statin therapy on cardiovascular outcomes in patients undergoing infrainguinal vascular surgery. Int J Cardiol 2005;104:264-8.

11. Kertai MD, Boersma E, Westerhout CM, van Domburg R, Klein J, Bax JJ, et al. Association between long-term statin use and mortality after successful abdominal aortic aneurysm surgery. Am J Med 2004;116:96-103.
12. Schouten O, Boersma E, Hoeks SE, Benner R, van Urk H, van Sambeek MR, et al. Fluvastatin and perioperative events in patients undergoing vascular surgery. N Engl J Med 2009;361:980-9.
13. Poldermans D, Bax JJ, Kertai MD, Krenning B, Westerhout CM, Schinkel AF, et al. Statins are associated with a reduced incidence of perioperative mortality in patients undergoing major noncardiac vascular surgery. Circulation 2003;107:1848-51.
14. O´Neil-Callahan K, Katsimaglis G, Tepper MR, Ryan J, Mosby C, Ioannidis JP, et al. Statins decrease perioperative cardiac complications in patients undergoing noncardiac vascular surgery: the Statins for Risk Reduction in Surgery (StaRRS) study. J Am Coll Cardiol 2005;45:336-42.
15. Christenson J. Preoperative lipid control with simvastatin reduces the risk for graft failure already 1 year after myocardial revascularization. Cardiovasc Surg 2001;9:33-43.
16. Abbruzzese TA, Havens J, Belkin M, Donaldson MC, Whittemore AD, Liao JK, et al. Statin therapy is associated with improved patency of autogenous infrainguinal bypass grafts. J Vasc Surg 2004;39:1178-85.
17. Henke PK, Blackburn S, Proctor MC, Stevens J, Mukherjee D, Rajagopalin S, et al. Patients undergoing infrainguinal bypass to treat atherosclerotic vascular disease are underprescribed cardioprotective medications: effect on graft patency, limb salvage, and mortality. Journal of Vascular Surgery 2004;39:357-65.
18. Hirsch AT, Haskal ZJ, Hertzer NR, Bakal CW, Creager MA, Halperin JL, et al. ACC/AHA 2005 Practice Guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on Practice Guidelines (Writing Committee to Develop Guidelines for the Management of Patients With Peripheral Arterial Disease): endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation; National Heart, Lung, and Blood Institute; Society for Vascular Nursing; TransAtlantic Inter-Society Consensus; and Vascular Disease Foundation. Circulation 2006;113:e463-654.
19. Hirsch AT, Criqui MH, Treat-Jacobson D, Regensteiner JG, Creager MA, Olin JW, et al. Peripheral arterial disease detection, awareness, and treatment in primary care. Jama 2001;286:1317-24.
20. McDermott MM, Mehta S, Ahn H, Greenland P. Atherosclerotic Risk Factors Are Less Intensively Treated in Patients with Peripheral Arterial Disease Than in Patients with Coronary Artery Disease. J Gen Intern Med 1997;12:209-15.
21. Mukherjee D, Lingam P, Chetcuti S, Grossman PM, Moscucci M, Luciano AE, et al. Missed opportunities to treat atherosclerosis in patients undergoing peripheral vascular interventions: insights from the University of Michigan Peripheral Vascular Disease Quality Improvement Initiative (PVD-QI2). Circulation 2002;106:1909-12.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.vascularqualityinitiative.org/wp-content/uploads/2016_PQRS_Information-v2-1.pdf

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

Attachment:

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

Attachment **Attachment:** LEB-defs-v.01.09_v1.doc

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

There have been no changes to the measure specifications.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*

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

Patients undergoing infrainguinal lower extremity bypass who are prescribed a statin medication at discharge.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*

Since hospitals have sufficient annual volume to generate accurate reporting levels, these are proposed for reporting every 12 months for hospital. Since surgeons have lower individual volume, we recommend annual reporting of the last 50 consecutive procedures, which may span more than one year, with suppression if < 10 procedures (ie, reported as too low volume to report).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

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

ANY registry that includes anatomic details or CPT procedure codes is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE)are examples of registries which capture detailed anatomic information, but the measure is not limited to these registries. It could also be used by other registries that capture this same information. No other registries are required for computation. Infrainguinal lower extremity bypass is defined as a bypass beginning at or below the external iliac artery and extending into the ipsilateral leg. It includes procedures with CPT codes 35656, 35556, 35583, 35666, 35566, 35585, 35671, 35571, 35587. The numerator is calculated as the number of patients age 18 and over undergoing such a procedure who are prescribed a statin medication at the time of discharge, which is also captured in the above registries.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*

All patients aged 18 years and older undergoing lower extremity bypass as defined above who are discharged alive, excluding those patients who are intolerant to statins.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*

Senior Care

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

ANY registry that includes anatomic details or CPT procedure codes is required to identify patients for denominator inclusion. The Society for Vascular Surgery Vascular Quality Initiative and the Vascular Study Group of New England are examples of registries that capture detailed anatomic information, but the measure is not limited to these registries. Infrainguinal lower extremity bypass is defined as a bypass beginning at or below the external iliac artery and extending into the ipsilateral leg. It includes procedures with CPT codes 35656, 35556, 35583, 35666, 35566, 35585, 35671, 35571, 35587. Only patients who are discharged alive are included in the denominator, and patients who are intolerant to statins are excluded, as described below.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

Chart documentation that patient was not an eligible candidate for statin therapy due to known drug intolerance, or patient died before discharge.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Chart documentation that patient was not an eligible candidate for statin therapy due to known drug intolerance, or patient died before discharge. These data are captured in the SVS VQI and VSGNE registries.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not required

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
NA

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
All patients age 18 and older undergoing infrainguinal LEB who were prescribed statin at discharge divided by (all patients over 18 undergoing infrainguinal LEB minus those intolerant to statins minus those who died before discharge).

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
NA

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
The Society for Vascular Surgery Vascular Quality Initiative Registry
The Vascular Study Group of New England Registry

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1519_MeasureTesting_MSF5.0_Data_v1.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1519 NQF Project: Surgery Endorsement Maintenance 2010

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)
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 field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

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)*:

A random sample of 100 patient records representing 5 procedures relevant to the measure from 5 different hospitals based on data collected during the past 2 years. In addition, in-hospital mortality was examined by claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007. These measures were originally tested in 2011 and this was the most recent data. All of the testing was approved by the NQF Steering Committee at the time that the measures were first approved in 2012. These measures are approved for PQRS reporting and working well. Regarding the sample and the data, this is an accepted testing practice to pull a sample for chart review to then compare to the data that was submitted to a registry.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
A nurse abstractor completed a form based on medical record review for the variables relevant to this measure. The results of this chart review were then compared with the original registry data. The Kappa statistic was used to judge reliability of the data. For mortality validation, claims data from each of 12 hospitals were matched to patient identified data within the VSGNE registry to compare discharge status (alive vs. dead). Any discrepencies were then further evaluated based on a medical record audit.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The key variables for this measure and testing results were:

1. Correct procedure (infrainguinal lower extremity bypass) performed. Kappa =1.0
2. Statin prescribed at discharge: Kappa=.80 (.11 SE)
3. Hospital mortality: Kappa = .91 (SE .01)

4. Age: 100% agreement, Kappa = 1.0 for age 18 or older categories.
5. Intolerant to statins: Kappa = 1.0

2b. VALIDITY. Validity, Testing, <u>including all Threats to Validity</u>: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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 reliability testing.

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
The validity testing of statin prescribed at discharge used the medical record as the gold standard. Discharge medications are routinely and carefully documented in both the discharge summary and discharge orders. The medication list on both the discharge summary and discharge orders were compared to confirm validity.

Patient age and hospital mortality have face validity. Correctness of operation type compared the operative report as the gold standard with the progress note in the medical record.

Data collected over time in VSGNE have been compared to published literature. Please see the evidence listed in the NQF form under importance.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
100% agreement was found between statin prescribed at discharge on the discharge summary and discharge orders. 100% agreement was also found between the procedure type reported in the operative note and that recorded in the daily progress notes.

Discharge statin use has been tracked in VSGNE for these procedures since 2003. Under a quality program, the proportion of patients discharged on statins has gradually improved, providing validity for this measurement.

POTENTIAL THREATS TO VALIDITY. (*<u>All</u> potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results*

demonstrating the need to specify them.)

2b3.1 Data/Sample for analysis of exclusions *(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)*:
2496 patients in the registry who underwent infrainguinal LEB between 2003-2010 in VSGNE, all patients in registry for this procedure

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
Rate determination

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
2% patients died in hospital
2% were alive but intolerant to statins
Of the remaining, 73% were discharged on statins.
Across 13 hospitals, the median statin prescribed at discharge rate was 73%, with an interquartile range of 69% to 80%. Across 63 individual providers, the median statin prescribed at discharge rate was 75%, with an interquartile range of 66% to 84%.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 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)*:
Not required for this process measure.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
NA

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
NA

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
NA

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
see section 1.b.3 and above 2,d,5

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
Standard statistial analysis to determine 95% confidence interval for hospitals and providers to determine practical difference from mean

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
see above

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
Other sources not available for testing.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
NA

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
NA

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: NA

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
Please see the new data under the importance sections of the NQF regular form per the requirement on the measure maintenance checklist.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition,
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
Yes

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
In the VSGNE experience which has been tracking statin usage since 2003, we have not experienced any difficulty with obtaining data related to statin usage. Our percent missing for perioperative statin use has been less than 2%. This has continued to be the case with the Vascular Quality Initiative (VQI) as well.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
SVS has never had a request to license any of our measures or set a fee structure.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Planned	Current Use (for current use provide URL)
	Payment Program Physician Quality Reporting System www.cms.hhs.gov

4a.1. **For each CURRENT use, checked above, provide:**

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

Physician Quality Reporting System (PQRS Measure Number 257) operated by the Centers for Medicare and Medicaid Services (CMS)

Physician Quality Reporting and Medicare payment adjustments.

PQRS is a national program.

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

As demonstrated by the data below, this measure say a level of improvement in 2010 when the measure was first created. It then had a slight reverse and since it has been included in PQRS and in the VQI as a qualified clinical data registry in 2015 there has again been improvement. Rate by year: 2010: 81.3%, 2011: 78.5%, 2012: 77.7%, 2013: 77.8%, 2014: 77.6% and 2015: 79.1%.

Also, we need to remember that the N=s for number of procedures captured in the VQI registry has also increased in each of these years, as well.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
It is possible to miss or inacurately code statin status. We have overcome this by providing each site with a list of generic and trade names for known statin medications.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
Related Measures: 0118 Antilipid therapy at discharge 0439 Discharged on statin medication

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Society for Vascular Surgery
Co.2 **Point of Contact:** Sarah, Murphy, SMurphy@vascularsociety.org, 312-334-2305-
Co.3 **Measure Developer if different from Measure Steward:** Society for Vascular Surgery
Co.4 **Point of Contact:** Jill, Rathbun, Jill_Rathbun@galileogrp.com, 703-217-7224-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
N/A

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

LOWER EXTREMITY BYPASS DEFINITIONS– v.01.09

If more than one response applies, select the most severe (highest number) response for each data field.

Pre-op Data

Smoking: Prior = quit ≥ 1 year ago. Current = still smoking within last 12 months. Include cigarettes, pipe, or cigar.

HTN (Hypertension): Defined as ≥ 140/90, either systolic or diastolic, at admission or within last 6 months, or clearly documented in medical record.

Beta-blockers: Peri-operative = started within one month before surgery or during surgery. Chronic = more than one month before surgery.

CAD Symptoms (Coronary artery disease**):** Stable angina = stable pattern or symptoms with or without antianginal medication. Unstable angina = new onset, increasing frequency, lasting > 20 min and/or rest angina.

CABG/PTCA: Coronary artery bypass, angioplasty, or stent.

CHF (Congestive Heart Failure): Documented CHF: Mild = SOB on exertion; Severe = SOB at rest, pulmonary edema, or pitting ankle edema. (Use 2 = mild if severity not documented.)

COPD: Not treated = COPD documented in record but not treated with medication. Medication includes theophylline, aminophylline, inhalers or steroids

Dialysis: Transplant = patient has functioning kidney transplant; Dialysis = currently on hemo- or peritoneal dialysis.

Creatinine: Last available measurement taken before procedure. If multiple measurements, use highest within 30 days of surgery.

Stress Test: Includes stress EKG, stress echo, nuclear stress scans, within 2 years of surgery.

Pre-admin living: Use last living status before any current, acute hospitalization, or rehab unit.

Previous Arterial:

Bypass - Any non-cardiac arterial bypass for occlusive disease
CEA - Carotid endarterectomy
Aneurysm Repair – Any known true arterial aneurysm repair (excluding cerebral or pseudo-aneurysm)
PTA/Stent – Of any non-cardiac artery
Major Amputation – Any amputation above the foot or hand

Pre-Op Medications: Taken within 36 hours of surgery. Statins include any HMG-CoA reductase inhibitor, such as Lipitor, Mevacor, Pravachol, Zocor, Lescol, etc. If Plavix is discontinued prior to surgery it should be coded = 0.

Pre-op Hemoglobin: Most recent pre-op hemoglobin within past 30 days.

Indication: Acute ischemia requires motor-sensory loss, sudden onset, and need for emergent treatment within 24 hours of presentation. Urgent = 12-72 hours. Emergent = <12 hours.

Pathology: If both aneurysm and occlusive disease, select the pathology that was the principal indication for the procedure.

Ambulation Pre-op: Chose best ambulation category experienced within one month of admission (lowest category).

Previous Ipsilateral/Contralateral: Inflow: aorto-iliac-femoral. Leg: intra-inguinal. Amputation: Major = above or below knee (loss of foot); Minor = within foot.

Pre-opABI, TBI: Use highest value from affected leg. TBI = toe-brachial index. Use actual units. Use 2.0 if non-compressible.

DSA/Angiogram: Digital subtraction or conventional arteriogram.

Procedure

Urgency: Urgent = required operation within 72 hours, but >12 hrs of admission. Emergent = required operation within 12 hrs of admission to prevent limb loss.

Recipient: Use most distal site if sequential bypass.

Vein type: Use composite for spliced vein from more than one vein site.

Concomitant Proximal Ipsilateral: Procedure performed proximal to or at origin of leg bypass graft to improve inflow during same operation.

Post-op Data

Wound infection: Culture positive or requiring antibiotic treatment.

Graft infection: Documented in record as exposed graft or graft infection.

Transfusion: Total of all PRBC transfusions pre-op, intra-op, and post-op during this hospitalization.

Myocardial Infarction: Troponin: by local standards for MI. EKG: new Q waves, new ST and T wave changes. Clinical: documentation of MI by clinical criteria or ECHO or other imaging modality.

Dysrhythmia: New rhythm disturbance requiring treatment with medications or cardioversion.

CHF: Pulmonary edema with requirement for monitoring or treatment in ICU.

Respiratory: Pneumonia = Lobar infiltrate on CXR and pure growth of recognized pathogen or 4+ growth of recognized pathogen in presence of mixed growth. Ventilator = required after initially extubated (if applicable).

Change renal function: New increase in creatinine of 0.5mg/dl. New dialysis includes peritoneal dialysis, hemodialysis, and hemo-filtration. (Applies to new dialysis not present pre-op.)

Bleeding; Infection; Thrombosis; Revision: Use 666 if Return to OR = 0.

Discharge patency: Primary = without other intervention; Primary-assisted = after intervention but without thrombosis; Secondary = after intervention for thrombosis.

Patency judged by: Use highest applicable modality. Palpable: clearly palpable pulse (not by Doppler). ABI: increase ABI (or TBI) ≥ 0.15 compared with pre-op.

Post-op ABI, TBI: Use highest value from affected leg. TBI = toe-brachial index. Use actual units. Use 2.0 if non-compressible.

Peri-operative Antibiotics: Use 0=no if antibiotic was not ordered. To use 1=yes, antibiotic must be ordered to be given within 1 hour prior to skin incision and must be ordered to be discontinued within 24 hrs of end of time of operation. To use 2=no for medical reason, a medical

reason must be documented in the chart that antibiotic not given. **Acceptable antibiotics include:** Ampicilin/sulbactam, Aztreonam, Cefazolin, Cefmetazole, Cefotetan, Cefuroxime, Ciprofloxacin, Clindamycin, Ertapenem, Erythromycin base, Gatifloxacin, Gentamicin, Levofloxacin, Metronidazole, Moxifloxacin, Neomycin, and Vancomycin.
1st-2nd Generation Cepahalosporin: (Cefazolin or Cefuroxime) Use response 1=yes, if ordered. If documented in medical record that not ordered for medical reason use 2. Otherwise use 0=no.

DENOMINATOR:

Patients who received an infra-inguinal lower extremity bypass

Denominator Criteria (Eligible Cases): All patients aged 18 years and older AND
Patient encounter during the reporting period (CPT): 35556, 35566, 35570, 35571, 35583, 35585, 35587, 35656, 35666, 35671

NUMERATOR:

Patients prescribed a statin medication at discharge

Numerator Options:

Performance Met:	Statin medication prescribed at discharge (G8816)
OR	
Denominator Exception: not	*Reason is documented in the medical record for why the* Statin therapy was prescribed **(G8815)**
OR	
Performance Not Met:	Statin therapy not prescribed at discharge, reason not given (G8817)

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1523
De.2. **Measure Title:** Rate of Open Repair of Abdominal Aortic Aneurysms (AAA) Where Patients Are Discharged Alive
Co.1.1. **Measure Steward:** Society for Vascular Surgery
De.3. **Brief Description of Measure:** Percentage of asymptomatic patients undergoing open repair of abdominal aortic aneurysms (AAA) who are discharged alive. This measure is proposed for both hospitals and individual providers.
1b.1. **Developer Rationale:** Elective AAA repair is offered to prolong life by avoiding AAA rupture, which is fatal in more than 85% of cases. Rupture risk is primarily assessed by AAA diameter, with larger AAAs more prone to rupture. Surgical treatment carries risk; however, of mortality and morbidity, which must be balanced against the risk of rupture in order to determine which patients will benefit from elective repair.

Based on the UK aneurysm trial, the accepted diameter threshold for elective AAA repair is 5.5 cm, although women have a slightly higher risk than men, so a threshold of 5 cm is usually recommended for women. The key concept of this proposed measure is that patients who are at low risk for AAA rupture (<6cm dia in men and <5.5 cm dia in women) should ONLY be offered elective AAA repair if their predicted operative mortality is low. This concept avoids the need for risk adjustment, since this is implicit in the decision to offer elective repair of AAAs. This measure will highlight variation in proper patient selection by reporting unadjusted mortality rates for surgery in patients with small AAAs in whom this rate should be universally low. Providers or hospitals with high mortality rates are either not performing safe surgery or are not properly selecting low risk patients. The measure specifically excludes patients with larger AAAs because risk adjustment would be needed for such cases, and accepted risk adjustment algorithms are not available.

S.4. **Numerator Statement:** Patients discharged alive/home following open repair of asymptomatic AAAs in men with < 6 cm diameter and women with < 5.5 cm diameter AAAs.
S.7. **Denominator Statement:** All elective open repairs of asymptomatic AAAs in men with < 6 cm dia and women with < 5.5 cm dia AAAs
S.10. **Denominator Exclusions:** = 6 cm minor diameter - men
= 5.5 cm minor diameter - women
Symptomatic AAAs that required urgent/emergent (non-elective) repair

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Facility

Original Endorsement Date: May 01, 2012 **Most Recent Endorsement Date:** May 01, 2012

IF this measure is paired/grouped, NQF#/title:
De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Submitted SVS measure: In-hospital mortality following elective endovascular repair of AAAs

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This outcome measure was initially endorsed in 2012 and calculates the percentage of asymptomatic patients undergoing open repair of abdominal aortic aneurysms (AAA) who are discharged alive.
- The UK aneurysm trial found that the accepted threshold for elective AAA repair is 5.5cm although women have a slightly higher risk of rupture. A threshold of 5cm is therefore recommended for women.
- The measure proposes that patients at low risk (<6cm in men, <5.5cm in women) should only be offered elective AAA repair if their predicted operative mortality is low. The developer states that there is no need for risk adjustment since this is implicit in decision to offer elective repair. Please refer to AAA definitions linked here.
- The measure will highlight variation in proper patient selection by reporting unadjusted mortality rates for patients with small AAA who should have lower mortality rates.
- Additional references were submitted for the Evidence in this submission.

Updates

The developer submitted updated evidence for this measure:

- Comprehensive Assessment of Factors Associated With In-Hospital Mortality After Elective Abdominal Aortic Aneurysm Repair.
  Hicks CW, Canner JK, Arhuidese I, Obeid T, Black JH 3rd, Malas MB.
  JAMA Surg. 2016 May 18. doi: 10.1001/jamasurg.2016.0782. [Epub ahead of print]

- The effect of hospital factors on mortality rates after abdominal aortic aneurysm repair.
  Dua A, Furlough CL, Ray H, Sharma S, Upchurch GR, Desai SS.
  J Vasc Surg. 2014 Dec;60(6):1446-51. doi: 10.1016/j.jvs.2014.08.111. Epub 2014 Oct 14.

- Surgeon case volume, not institution case volume, is the primary determinant of in-hospital mortality after elective open abdominal aortic aneurysm repair.
  McPhee JT, Robinson WP 3rd, Eslami MH, Arous EJ, Messina LM, Schanzer A.
  J Vasc Surg. 2011 Mar;53(3):591-599.e2. doi: 10.1016/j.jvs.2010.09.063. Epub 2010 Dec 8.

If the developer provided updated evidence for this measure:

- *Does the stated rationale support the relationship of the health outcome to processes or structures of care?*

Guidance from the Evidence Algorithm

Assessment of performance of health outcome (Box 1) → relationship between health outcome and action supported by rationale → Pass

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

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

Maintenance measures – increased emphasis on gap and variation

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

- Using data from the Dartmouth-CMS-FDA Collaborative, the developer reports that in 27 hospital referral regions, rates of AAA repair were at least 30% higher than the United States average of 1.0 per 1,000 Medicare enrollees. In 44 hospital referral regions, rates were more than 25% lower than the national average.
- The developer includes information from the original submission that VSGNE data show that among 12 centers and 55 providers treating 1,289 patients with small AAA, the median mortality rate for both men and women was 0%, and ranged from 0% to 10%.
- The developer notes that in 2016, the VQI registry was reviewed for the time period 2010-2015. Of 170 centers, the discharged alive percentage ranged from 100% to 95%. Of all the cases (n=4,266) there was a mortality rate of 3.3%.

Disparities

- In the review of 2010-2015 VQI registry data, the developer reported that for those aged 80 and older the rate of mortality was 7.4% versus those age 60 and below whose rate was 0.4%. Developers also report a higher rate of mortality in females and non-Hispanics although actual rates were not provided.

Questions for the Committee:

- *Is there a sufficient gap in care that warrants a national performance measure?*
- *Is there expected variation in performance if reported at the physician level versus at the facility level?*
- *Are you aware of evidence that disparities exist in this area of healthcare or that such data would exist in registry?*

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

Committee pre-evaluation comments

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

1a.

- This study's end point is unadjusted in hospital mortality in asymptomatic patients undergoing elective open resections of AAAs which are <6cms in diameter in men and <5.5cms in women. The evidence supports that this is reasonable to do as long as the operative risk is low. The objective of the study is to determine if it is indeed low, although low is not defined. There is evidence that the proposers identify that supports this recommendation.
- I believe that the evidence supports the measure focus and recommend a rating of pass.
- Greater validity of AAA size than hospital mortality
- Suppressing docs < 10 cases -- why? Wouldn't this group show the greatest potential benefit?
- Low mortality overall, so ability to identify outliers will be very low.
- I like this measure for public accountability, not for quality improvement
- What about the anesthesiologist?
- This is an outcome measure. The relationship between the measured outcome and at least one healthcare action is clear and supported by the rationale.

1b.

- When the model was developed the median mortality rate for males and females was zero and varied from 0% to 10%, representing room for improvement across sites. Over the period of 2010-2015 with 170 centers participating the percentage discharged alive varied from 95% to 100% with a mortality of 3.3%. There are studies nationally showing disparity in mortality across regions of the US which perhaps could be narrowed by collecting this data.
- In regard to the question of whether the gap warrants approval of this measure, the geographic differences would support that although the gap is small. The gap has also decreased since the approval of the measure which is a desired outcome.
- In regard to whether there is expected variation in performance if reported at the physician level vs facility level, it would be expected at both levels but because of having adequate numbers of patients would require lumping several years of data for the individual surgeons.
- In regard to disparities, as noted above, they are present regionally.

- I would rate the opportunity for improvement as low but significant, because in this asymptomatic ""low risk"" group of patients it should be close to zero.
- There is overall good performance on this measure, however the shift from open to endovascular procedures overall is reducing the number of open cases. The developers demonstrate that low volume providers have poorer outcomes. I would rate this as moderate.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- **Data source(s):** Electronic Clinical data: registry. Registries used for this measure are The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE). VQI and VSGNE include hospitalization details and symptom status.

Specifications:

- This measure is specified at the Clinician: Group/Practice, Clinician: Individual facility level in the hospital/acute care facility.
- The numerator includes patients discharged alive/home following open repair of asymptomatic AAA in men with <6cm diameter and women with <5.5 cm diameter AAAs.
- The denominator includes all elective open repairs of asymptomatic AAAs in men with <6cm diameter and women with <5.5 cm diameter AAAs.
- Patients are excluded if the AAA diameter is equal to or above 6cm in men and 5.5 cm in women. Symptomatic AAAs that require repair (non-elective) are also excluded.
- The developers indicate that as part of acceptance into the Centers for Medicare & Medicaid Services Physician Quality Reporting System, the measure needed to be framed in the positive (i.e., discharged alive) versus the negative (i.e., mortality).
- The measure is calculated with the number of deaths divided by the number of cases.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate data elements and definitions included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented using the registry and outside the registry?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Testing has not been updated since the previous endorsement although the measure has now been framed as patients discharged alive. A summary of testing is provided below.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- A random sample of 100 patient records were reviewed, representing 5 AAA procedures from 5 different

hospitals collected during the "past 2 years". (Note that the 'past 2 years' refers to the time period prior to when the measure was first endorsed)

- Hospital reporting is proposed for every 12 months based on sufficient volume.
- Annual reporting of the last 50 consecutive procedures for surgeons (which may span more than one year) with suppression of <10 procedures is recommended.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007.
- Chart abstraction was completed with results related to relevant variables compared to registry data.
- Developers analyzed the level of agreement between the chart and registry data using the Kappa statistic.
- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched to patient data within the registry to compare discharge status (alive vs dead).
- Any discrepancies were further evaluated based on a medical record audit.

Results of reliability testing

- Data element validity testing was used to support the reliability of the measure.
- Kappa statistics indicated strong agreement for identification of the correct procedure (AAA) performed (1.0), AAA diameter (1.0), elective repair (1.0) and hospital mortality (.91).
- The developer notes there was no significant difference in AAA mean diameter between registry (56.7 mm) and chart audit (56.6mm).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Are the method, samples, and outcomes of testing clear?*
- *Is testing for in hospital mortality sufficient for how the measure is now framed (patients discharged alive)?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) →Patient level data validity (Box 3) → (Box 10 of validity algorithm) →Appropriate method to assess data elements (Box 11) → Moderate certainty that data used in the measure are valid (Box 12a) →Highest possible rating is moderate.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- Testing has not been updated since the previous endorsement although the measure has now been framed as patients discharged alive. A summary of testing is provided below.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- A random sample of 100 patient records was reviewed, representing 5 AAA procedures from 5 different hospitals.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007.
- Chart abstraction was completed with results compared to registry data.
- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched to patient data within the registry to compare discharge status (alive vs dead).

Results of validity testing

- Kappa statistics indicated strong agreement, at the hospital level, for identification of the correct procedure (AAA) performed (1.0), AAA diameter (1.0), elective repair (1.0) and hospital mortality (.91).
- Clinician-level information is not provided.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*
- *Are the method, samples, and outcomes of testing clear?*
- *Is testing for in hospital mortality sufficient given how the measure is now framed (patients discharged alive)?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients are excluded from the measure based on diameter of the AAA and non-elective repair.
- Exclusions analysis was completed for 1,201 patients undergoing elective AAA repair in VSGNE.
- Data were from 2003 to 2010 and included 886 men and 315 women
- AAAs were analyzed with 6cm diameter cut point in men and 5.5 cm diameter cut point in women .
- Numbers for the subsets below not provided
- Data for men from 10 centers showed
  - median mortality of 0% for those with <6cm AAA; range, 0 to 4.1%;
  - median mortality of 0% for those with >=6cm AAA; range, 0 to 10.4%.
- Data for women from 9 centers showed
  - median mortality of 0% for those with <5.5 cm AAA; range, 0% to 10%
  - median mortality of 1.1% for those with >=5.5 cm AAA; range, 0% to 20%.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included? ☐ **Yes** ☒ **No**

Risk adjustment summary:

- The measure is not risk adjusted.
- The developer gave rationale that "risk adjustment is complex for AAA repair, and accepted algorithms do not yet exist". The developer provided a list of open infrarenal AAA definitions that includes a number of factors for

pre-operative consideration.

- In addressing disparities, the developer states that such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region. The developer also reported higher mortality for patients 80+ years of age (7.4%) vs those 60 and below (0.4%) as well as higher mortality rate in females and non-Hispanics.

Questions for the Committee:

- *Do you agree with the developer's rationale regarding risk adjustment?*
- *What is the Committee's expectation regarding consideration of SDS factors in maintenance measures?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer reports using standard statistical analysis to determine 95% confidence interval for hospital and providers
- Results outlined in 2b3 above show that in men,
  - median mortality of 0% for those with <6cm AAA; range, 0 to 4.1%;
  - median mortality of 0% for those with >=6cm AAA; range, 0 to 10.4%.
- Data for women from 9 centers showed
  - median mortality of 0% for those with <5.5 cm AAA; range, 0% to 10%
  - median mortality of 1.1% for those with >=5.5 cm AAA; range, 0% to 20%.

Question for the Committee:

- *Is the Committee comfortable with using mortality data to interpret "discharge live" for the measure?*
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- The developer states no other data sources are available.

2b7. Missing Data

- The developer reports less than 1% missing data.

Guidance from the Validity Algorithm

Precise specifications (Box 1) → All threats to validity assessed (Box 2) →Insufficient

Potential threats to validity around risk adjustment and SDS factors result in preliminary rating.

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

Committee pre-evaluation comments

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

2a1.

- First, this very importantly has clinical data obtained from electronic patient records which captures clinical details including symptoms, radiology findings, etc. The definitions of the patient population are clearly defined in terms of aneurysm size, lack of symptoms, elective status, and the exclusion of patients is evidence bases in terms of those with aneurysms larger than 6cm in men and 5.5 in women being excluded because of the increased risk of adverse outcomes if they are followed rather than resected, assuming their operative risk is not prohibitive. Discharged alive is a very clear outcome and easily defined. The logic is clear and it is likely that this measure can be consistently implemented and it can be duplicated outside of the registry.
- I believe that the measure could be improved by developing a risk model for mortality or survival, but based on their data the patients entered appear to be low risk based on their outcome.
- Reliability testing was excellent with Kappa statistics indicating strong agreement for identification of correct procedure, AAA diameter, elective repair and hospital mortality.
- I believe that the sample size was adequate, the methods clear, that hospital mortality is easily translated to discharged alive, and that if this was used at the clinician level it would be reliable if there is an adequate

sample size which probably would required procedures over several years.
- I would rate the reliability as high.

2a2.
- I believe that the number tested was adequate for this, particularly when considering the consistently high Kappa statistics and the simplicity of the measure. The testing was done at the level of the group or institution, not the individual clinician, but I see no reason why it could not be applied to the clinician if the volume of cases was enough to draw meaningful conclusions from the findings.
- The calculation is clear.

2b1.
- I find the specifications to be consistent with the evidence.
- Not updated. Agree with the developer's risk adjustment rationale.
- I question why the measure is specific at a threshold of 5.5 cm for women when the rationale states that a threshold of 5 cm is the usual recommendation for women.

2b2.
- The testing was performed on a sample of 100 patient records representing 5 AAA procedures from 5 hospitals. Mortality was validated by comparing to claims data from 12 hospitals participating in the VSGNE Registry matched to patient data. Kappa statistics indicated strong agreement across all of the main parameters.
- No additional testing was performed. The testing was adequate at both the institution and the provider level.

2b3.
- Since this measure is by definition confined to asymptomatic patients with clearly defined by size aneurysms who are in the opinion of the clinicians to be at low risk for surgery, it is totally appropriate to exclude symptomatic patients with larger aneurysms. Presumably, there are or will be other measures to capture that population of patients.
- In regards to risk adjustment, in my humble opinion it would be advantageous for the SVS to have a risk model that would be appropriate for all patients undergoing open repair of AAAs. Although that is not absolutely essential for the measure, I believe that it would strengthen it.
- In regard to meaningful differences, I do not believe the developers addressed this specifically, but in this population if only 95% or less patients left the hospital alive that finding would potentially draw attention to issues with quality of care.
- No performance scores were compared. this measure only captures patient percentage discharged from the hospital alive. Individual sites would then be able to judge their performance as compared to their peers, ie, better or worse or the same as.
- We were not made aware of any significant missing data but this should be asked at the Committee meeting.
- Overall, I would rate validity as moderate.
- No

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data are coded by someone other than the person obtaining original information.
- The developer reports that VSGNE has been tracking hospital mortality since 2003 and they have not had difficulty obtaining mortality data. The developer reports a percent missing of less than 1%.
- There are no licenses or costs associated with the use of this measure.

Questions for the Committee:
- o *Are the required data elements routinely generated and used during care delivery?*
- o *Are there fees to belong to the registry?*

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

Committee pre-evaluation comments
Criteria 3: Feasibility

- This measure appears to be very feasible and relatively simple to utilize. The data burden seems reasonable, straight forward and doable.
- All elements are present in the vascular surgery registries which do not require a fee. Data abstracting and reporting does require resources.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure

- Centers for Medicare and Medicaid Services, Physician Quality Reporting Services (measure 417)

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- PQRS is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). PQRS measures are used for public reporting on the Physician Compare website and for the quality component of the Value-Based Payment Modifier (VBPM).

Improvement results

- The developer reports that as the sample size in the VQI databased increased for this measure, so did the rate of mortality (in 2011 and 2012); as the number of cases has stabilized, there was a reduction in mortality in 2013 and to 2.8% in 2015.
- In the 2016 review of VQI registry data for data collected from 2010-2015, the discharged alive percentage ranged from 100% to 95%. Of all 4,266 cases reported, the developers report a mortality rate of 3.3%.
- Although the measure is reported in PQRS, these data were not provided by the developer

Unexpected findings (positive or negative) during implementation

- The developer reports that size measurements of AAA should not significantly impact the measure and that symptom status is easily validated during chart review.

Potential harms

- The developer does not note any potential harms.

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- The measure is being publically reported through PQRS to the Physician Compare Website for quality reporting for the Value-Based Payment Modifier.
- I would rate usability as high.
- Publicly reported - used in PQRS.

Criterion 5: Related and Competing Measures
Related or competing measures • N/A Harmonization • N/A

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1523 NQF Project: Surgery Endorsement Maintenance 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT
Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence. *Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria.* (evaluation criteria)
1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*: discussed above

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:
The endpoint of inhospital mortality is the accepted primary endpoint for both elective AAA repair. Variation in outcome has been established in randomized trials,cohort studies and meta analyses. This outcome measure has face validity among all providers of this service. Studies cited above have shown substantial variation in outcomes by provider when elective AAA repair is performed in patients with AAAs.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*:

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Expert opinion.

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Motality is the reporting standard recommended by the Society for Vascular Surgery, and has been used in multiple RCTs.

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:
Fillinger M. (2010) Abdominal Aortic Aneurysms: Evaluation and Decision Making. In J. Cronenewett & KW. Johnston (Eds.), Rutherford´s Vascular Surgery (1928-1948) Saunders Elsevier. Philadelphia.

2016

Comprehensive Assessment of Factors Associated With In-Hospital Mortality After Elective Abdominal Aortic Aneurysm Repair.
Hicks CW, Canner JK, Arhuidese I, Obeid T, Black JH 3rd, Malas MB.
JAMA Surg. 2016 May 18. doi: 10.1001/jamasurg.2016.0782. [Epub ahead of print]

The effect of hospital factors on mortality rates after abdominal aortic aneurysm repair.
Dua A, Furlough CL, Ray H, Sharma S, Upchurch GR, Desai SS.
J Vasc Surg. 2014 Dec;60(6):1446-51. doi: 10.1016/j.jvs.2014.08.111. Epub 2014 Oct 14.

Surgeon case volume, not institution case volume, is the primary determinant of in-hospital mortality after elective open abdominal aortic aneurysm repair.
McPhee JT, Robinson WP 3rd, Eslami MH, Arous EJ, Messina LM, Schanzer A.
J Vasc Surg. 2011 Mar;53(3):591-599.e2. doi: 10.1016/j.jvs.2010.09.063. Epub 2010 Dec 8.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
None

1c.17 Clinical Practice Guideline Citation: None

1c.18 National Guideline Clearinghouse or other URL: None

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: N/A

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: N/A

1c.24 Rationale for Using this Guideline Over Others: Mortality is the accepted endpoint used in all trials. Restricting the AAA risk by confining the analysis to small or moderate AAAs is explained above.

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: 1c.26 Quality: 1c.27 Consistency:

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1523_Evidence_MSF5.0_Data_2016.doc

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Elective AAA repair is offered to prolong life by avoiding AAA rupture, which is fatal in more than 85% of cases. Rupture risk is primarily assessed by AAA diameter, with larger AAAs more prone to rupture. Surgical treatment carries risk; however, of mortality and morbidity, which must be balanced against the risk of rupture in order to determine which patients will benefit from elective repair.

Based on the UK aneurysm trial, the accepted diameter threshold for elective AAA repair is 5.5 cm, although women have a slightly higher risk than men, so a threshold of 5 cm is usually recommended for women. The key concept of this proposed measure is that patients who are at low risk for AAA rupture (<6cm dia in men and <5.5 cm dia in women) should ONLY be offered elective AAA

repair if their predicted operative mortality is low. This concept avoids the need for risk adjustment, since this is implicit in the decision to offer elective repair of AAAs. This measure will highlight variation in proper patient selection by reporting unadjusted mortality rates for surgery in patients with small AAAs in whom this rate should be universally low. Providers or hospitals with high mortality rates are either not performing safe surgery or are not properly selecting low risk patients. The measure specifically excludes patients with larger AAAs because risk adjustment would be needed for such cases, and accepted risk adjustment algorithms are not available.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
There is significant regional variation in rates of open AAA repair, indicating a performance gap. In 27 hospital referral regions, rates of AAA repair were at least 30% higher than the United States average of 1.0 per 1,000 Medicare enrollees. In 44 hospital referral regions, rates were more than 25% lower than the national average.(1)

Where these data have been monitored and reported to providers in VSGNE since 2003, among 12 centers and 55 providers treating 1289 patients with small AAAs the median mortality rate for men and women with small AAAs as defined above is 0%, but the range is 0-10%, indicating both a performance gap and opportunity for further improvement.

In 2016, the VQI Registry conducted a study specifically for NQF measure maintenance over the time period of 2010 - 2015. Of the 170 centers in the study, the discharged alive percentage ranged from 100% to 95%. Overall of all the cases - 4,266 - reported to the registry over the five year period, there was a morality rate of 3.3% demonstrating a continued need for improvement.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
(1)Dartmouth-CMS-FDA Collaborative, "Trends and Regional Variation in Abdmonial Aortic Anweurysm Repair, February 1, 2006.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region.

In this study conducted in 2016 for NQF measure maintenance, the data showed that for those aged 80+ the rate of mortality was 7.4% versus those age 60 and below whose rate was 0.4%. It also showed a higher rate of mortality in females and non-hispanics.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
not available

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare.**

List citations in 1c.4.
An international population-based study found that an aneurysm was present in 8.9% of men and 2.2% women (p < 0.001).(1)
In the United States, ruptured AAAs are the 15th leading casue of death overall and the 10th leading casue of death in males over 55 years, a rate than has held steady for the past 2 decades. (2)
Ruptured aneurysms are fatal in about 80% of cases. (3)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
(1) Singh K et al. Am. J. Epidemiol. (2001) 154 (3): 236-244.
(2) Fillinger M. (2010) Abdominal Aortic Aneurysms: Evaluation and Decision Making. In J. Cronenewett & KW. Johnston (Eds.), Rutherford´s Vascular Surgery (1928-1948) Saunders Elsevier. Philadelphia.
(3) May J, White GH, Stephen MS, Harris JP. J Vasc Surg. 2004 Nov;40(5):860-6.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.vascularqualityinitiative.org/wp-content/uploads/2016_PQRS_Information-v2-1.pdf

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** LEB-defs-v.01.09_v1-636009094258447860.doc

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
This measure was accepted for PQRS starting with the 2016 reporting year. As part of that process, CMS asked that the measure be put in the positive - i.e. discharged alive - versus the negative of mortality.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients discharged alive/home following open repair of asymptomatic AAAs in men with < 6 cm diameter and women with < 5.5 cm diameter AAAs.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Since hospitals have sufficient annual volume to generate accurate reporting levels, these are proposed for reporting every 12 months for hospital. Since surgeons have lower individual volume, we recommend annual reporting of the last 50 consecutive procedures, which may span more than one year, with suppression if < 10 procedures (ie, reported as too low volume to report).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
ANY registry that includes hospitalization details, AAA diameter and discharge status is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Any registry that collects this data could report on this measure. Patients who died in hospital following elective open infrarenal AAA repair if their aneurysm was asymptomatic (< 6cm dia in men, <5.5 cm dia in women, judged by preoperative imaging (CT, MR or ultrasound)).

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All elective open repairs of asymptomatic AAAs in men with < 6 cm dia and women with < 5.5 cm dia AAAs

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
ANY registry that includes hospitalization details, AAA diameter and discharge status is required to identify patients for denominator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Patients who underwent elective open AAA repair are included if their aneurysm was asymptomatic (< 6cm dia in men, <5.5 cm dia in women, judged by preoperative imaging(CT, MR or ultrasound)).

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
= 6 cm minor diameter - men
= 5.5 cm minor diameter - women
Symptomatic AAAs that required urgent/emergent (non-elective) repair

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Patients undergoing non-elective open repair of symptomatic AAAs or those with AAAs larger than the diameters noted above.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not required

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See "Scientific Acceptablility" section for rationale

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Identify denominator, exclude non-elective repair of symptomatic or ruptured patients and men with AAA >6 cm, and women with AAA >5.5, find number of deaths
Outcome = deaths/ # cases

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Society for Vascular Surgery Vascular Quality Initiative Registry
Vascular Study Group of New England Registry

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1523_MeasureTesting_MSF5.0_Data_v1.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1523 NQF Project: Surgery Endorsement Maintenance 2010

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)

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 field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability*.)

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)*:

A random sample of 100 patient records representing 5 procedures relevant to the measure from 5 different hospitals based on data collected during the past 2 years. In addition, in-hospital mortality was examined by claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007. These measures were originally tested in 2011 and this was the most recent data. All of the testing was approved by the NQF Steering Committee at the time that the measures were first approved in 2012. These measures are approved for PQRS reporting and working well. Regarding the sample and the data, this is an accepted testing practice to pull a sample for chart review to then compare to the data that was submitted to a registry.

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

A nurse abstractor completed a form based on medical record review for the variables relevant to this measure. The results of this chart review were then compared with the original registry data. The Kappa statistic was used to judge reliability of the data. For mortality validation, claims data from each of 12 hospitals were matched to patient identified data within the VSGNE registry to compare discharge status (alive vs. dead). Any discrepencies were then further evaluated based on a medical record audit.

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

The key variables for this measure and testing results were:

1. Correct procedure (open infrarenal AAA repair) performed. Kappa =1.0

2. AAA diameter: Based on 60 measurement, the mean diameter was 56.7 mm in the registry, 56.6 mm in the chart audit, no significant difference. Further, in on cases was the category of size based on the cut points of 6 cm in men and 5.5 cm in women different, Kappa = 1.0 for these categories.

3. Hospital mortality: Kappa = .91 (SE .01)

4. Elective(vs urgent or emergent); Kappa=1.0

Category II codes were created and they are included in the specification section of the main application.

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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 reliability testing

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:

comparison of rates with published literature. Please see the evidence listed in the NQF form under importance.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:

In VSGNE, in hospital mortality for open AAA repair is 4-8%, and shows appropriate variation among hospitals, using this measure. This corresponds well to the published literature for elective AAA repair.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(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)*:

1201 patients undergoing open elective AAA repair in VSGNE, all patients (ie, all AAA diameters treated), 2003-2010. 886 men, 315 women

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:

rate calculation based on AAA dia size. AAAs were analyzed with 6 cm dia cutpoint in men and a 5.5 cm dia cutpoint in women, as described below.

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*

Men, < 6cm AAA, mdn 0% mortality, range 0-4.1% among 10 centers

Men, >= 6 cm dia, mdn 0% mortality, range 0-10.4% among 10 centers

Women, < 5.5 cm dia AAAs, mdn mortality 0%, range 0-10% among 9 centers

Women, >= 5.5 cm dia AAAs, mdn mortality 1.1%, range 0-20% among 9 centers. These results allowed for the tailoring of the measure to the appropriate patients.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 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)*:

This measure was designed to avoid the need for risk adjustment, because risk adjustment is complex for AAA repair, and accepted algorithms do not yet exist. In patients with AAAs, with low rupture risk, it is incumbent on the surgeon to factor in the risk-benefit of elective, prophylactic repair, since a high operative mortality will eliminate any benefit of AAA repair. Women have higher rupture risk than men, so by focusing this measure on AAAs < 5.5 cm in women and < 6 cm in men, the non-risk-adjusted mortality is a fair comparison of surgical outcome in the opinion of the sponsor, the Society for Vascular Surgery, and it represents a very important outcome to measure.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

N/A

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

N/A

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment: N/A

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

see section 1.b.3 and above 2,d,5

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:

Standard statistial analysis to determine 95% confidence interval for hospitals and providers to determine practical difference from mean

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:

no other data sources available

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:

2c. Disparities in Care: H☐ M☐ L☐ I☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: NA

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:

Disparities have not been reported. Please see the new data under the importance sections of the NQF regular form per the requirement on the measure maintenance checklist.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?

(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐

Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition,
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

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

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

Yes

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

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

In the VSGNE experience which has been tracking hosptital mortality as a major endpoint since 2003, we have not experienced any difficulty with obtaining data related to this endpoint. Our percent missing for this variable has been less than 1%.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
SVS has never been asked to license any measures and has not established a fee structure.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Payment Program PQRS www.cms.hhs.gov

4a.1. **For each CURRENT use, checked above, provide:**

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

This measure was approved for use in the Physician Quality Reporting Program as of January 1,2016 as measure number 417. This program is run by the Centers for Medicare and Medicaid Services and it is a national program

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

As the sample size for this measure in the VQI database increased, the rate of mortality increased in 2011 and 2012. However, as the the number of cases per year stablized, we saw a reduction in 2013 and then a reduction again in 2015 to 2.8%.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of**

initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

Size measurements of AAA should not significantly impact the measure, and symptom status is easily validated during chart review. We have not found inaccuracy in this measure.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

No

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

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

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

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

Are the measure specifications completely harmonized?

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

5b. Competing Measures

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

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

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

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Society for Vascular Surgery
Co.2 **Point of Contact:** Sarah, Murphy, SMurphy@vascularsociety.org, 312-334-2305-
Co.3 **Measure Developer if different from Measure Steward:** Society for Vascular Surgery
Co.4 **Point of Contact:** Jill, Rathbun, Jill_Rathbun@galileogrp.com, 703-217-7224-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

OPEN INFRARENAL AAA DEFINITIONS – v.01.09

If more than one response applies, select the most severe (highest number) response for each data field.

Pre-op

Smoking: Prior = quit ≥ 1 year ago. Current = still smoking within last 12 months. Include cigarettes, pipe, or cigar.
HTN (Hypertension): Defined as ≥ 140/90, either systolic or diastolic, at admission or within last 6 months, or clearly documented in medical record.
Beta-blockers: Peri-operative = started within one month before surgery or during surgery. Chronic = more than one month before surgery.
CAD Symptoms (Coronary artery disease**):** Stable angina = stable pattern or symptoms with or without anti-anginal medication. Unstable angina = new onset, increasing frequency, lasting > 20 min and/or rest angina.
CABG/PTCA: Coronary artery bypass, angioplasty, or stent.
CHF (Congestive Heart Failure): Documented CHF: Mild = SOB on exertion; Severe = SOB at rest, pulmonary edema, or pitting ankle edema. (Use 2 = mild if severity not documented.)
COPD: Not treated = COPD documented in record but not treated with medication. Meds include theophylline, aminophylline, inhalers or steroids
Dialysis: Transplant = patient has functioning kidney transplant; Dialysis = currently on hemo- or peritoneal dialysis.
Creatinine: Last available measurement taken before procedure. If multiple measurements, use highest within 30 days of surgery.
Stress Test: Includes stress EKG, stress echo, nuclear stress scans, within 2 years of surgery.
Pre-admin living: Use last living status before any current, acute hospitalization or rehab unit.
Previous Arterial:
Bypass - Any non-cardiac arterial bypass for occlusive disease
CEA - Carotid endarterectomy
Aneurysm Repair – Any known true arterial aneurysm repair (excluding cerebral or pseudo-aneurysm)
PTA/Stent – Of any non-cardiac artery
Major Amputation – Any amputation above the foot or hand
Pre-Op Medications: Taken within 36 hours of surgery. Statins include any HMG-CoA reductase inhibitor, such as Lipitor, Mevacor, Pravachol, Zocor, Lescol, etc. If Plavix is discontinued prior to surgery it should be coded = 0.
Pre-op Hemoglobin: Most recent pre-op hemoglobin within past 30 days.

Family history of AAA: First-degree relative (parents, sibling, aunt, uncle, child)
Prior Aortic Surgery: AAA = infrarenal aneurysm repair. SAAA = Suprarenal aneurysm repair. Bypass = A-1 or A-F for occlusive disease. Other = endarterectomy or other.
Ejection Fraction: Left ventricular ejection fraction (%), by Echo, nuclear scan, or cath estimate, within 6 months
Maximum AP AAA diameter: Largest AP diameter. If AP not specified, use largest diameter. If multiple imaging modalities, use most accurate in following hierarchy: CT>MRI>Echo>arteriogram.

Iliac aneurysm: Iliac diameter > 1.5 cm. Maximum diameter of largest iliac artery, common, or internal.

Procedure

Urgency: Symptomatic = surgery within 24 hours of pain and/or tenderness without rupture. Ruptured = diagnosis at operation.
Conversion from endovascular: Early = within 30 days, late = >30 days
Renal/visceral ishcemic time: Include any aortic re-clamp time for hypotension.
Exposure: Anterior = transperitoneal
Distal anastomosis: Most distal extent of either right or left limb if bifurcated.
Graft Diameter: Body size = diameter of most proximal portion of graft.
Total procedure time: From incision to closure.

Concomitant Procedure

Thromboembolectomy: For inadequate limb perfusion after initial completion of distal anastomosis via Fogarty or extension of graft (bypass).

Ruptured AAA Repairs Only

Lowest pre-intubation BP: After arrival at hospital (lowest prior to intubation). Use systolic pressure.
Mental status: Normal alert and oriented; Disoriented to person, place, or time.
Delayed closure: Fascia not closed at initial operation to avoid compartment syndrome.

Post-op Data

Time to extubation: In OR; otherwise, beginning upon departure from OR
Vasopressor Required Post-Op: Dopamine≥5mcg/kg/min, or neosynepherine, levophed, epinepherine, vasopressin, or other IV vasopressor during hospitalization.
ICU stay: Any portion of 24 hours = 1 day.
Transfusion: Total of all PRBC transfusions pre-op, intra-op, and post-op during this hospitalization.
Myocardial Infarction: Troponin: by local standards for MI. EKG: new Q waves, new ST and T wave changes. Clinical: documentation of MI by clinical criteria or ECHO or other imaging modality.
Dysrhythmia: New rhythm disturbance requiring treatment with medications or cardioversion.
CHF: Pulmonary edema with requirement for monitoring or treatment in ICU.
Respiratory: Pneumonia = Lobar infiltrate on CXR and pure growth of recognized pathogen or 4+ growth of recognized pathogen in presence of mixed growth. Ventilator = required after initially extubated (if applicable).
Change renal function: New increase in creatinine of 0.5mg/dl. New dialysis includes peritoneal dialysis, hemodialysis, and hemo-filtration. (Applies to dialysis only if not required pre-op.)
Leg ischemia/emboli: Loss of previously palpable pulses, loss of previously present Doppler signals, decrease of >0.15 in ABI, or blue toe.
Bowel ischemia: Diagnosed by colonoscopic evidence of ischemia, bloody stools in a patient who dies prior to colonoscopy or laparotomy, or presumptive diagnosis with conservative treatment.
Peri-operative Antibiotics: Use 0=no if antibiotic was not ordered. To use 1=yes, antibiotic must be ordered to be given within 1 hour prior to skin incision and must be ordered to be discontinued within 24 hrs of end of time of operation. To use 2=no for medical reason, a medical reason must be documented in the chart that antibiotic not given. **Acceptable antibiotics include:** Ampicilin/sulbactam, Aztreonam, Cefazolin, Cefmetazole, Cefotetan, Cefuroxime, Ciprofloxacin, Clindamycin, Ertapenem, Erythromycin base, Gatifloxacin, Gentamicin, Levofloxacin, Metronidazole, Moxifloxacin, Neomycin, and Vancomycin.
1st-2nd Generation Cepahalosporin: (Cefazolin or Cefuroxime) Use response 1=yes, if ordered. If documented in medical record that not ordered for medical reason use 2. Otherwise use 0=no.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1540
De.2. **Measure Title:** Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Endarterectomy
Co.1.1. **Measure Steward:** Society for Vascular Surgery
De.3. **Brief Description of Measure:** Percentage of patients age 18 or older without carotid territory neurologic or retinal symptoms within the one year immediately preceding carotid endarterectomy (CEA) who experience stroke or death following surgery while in the hospital. This measure is proposed for both hospitals and individual surgeons.
1b.1. **Developer Rationale:** Numerous manuscripts have noted variation in the combined endpoint of stroke or death following carotid endarterectomy. In the Medicare population, the outcome has been shown to vary substantially as a function of hospital volume. This is an important consideration, since it is widely recognized that many surgeons and centers performing CEAs do not meet the high standards of the randomized trials which established the benefit of such treatment. It has been shown that mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and fully 2.5% in low volume hospitals (Ref 6). Given that the stroke rate is generally 3 times the mortality rate, this suggests that some centers/surgeons are not achieving optimal results. A recent survey in Canada found that 45% of hospitals are not meeting published guidelines (Ref 7). Adoption of this outcome measure in the United States would likely disclose similar results and lead to quality improvement when this information was provided to surgeons and centers. This effect has been demonstrated in a midwest regional study by Kresowik et al where stroke and death rate after CEA improved after providing outcome data (Ref 5). The VSGNNE has shown that regional results are good for CEA outcomes, but significant variation does exist between surgeons and centers (Ref 8). Postoperative stroke or death is the accepted outcome paramenter for this surgery, and its measurement and reporting would demonstrate variation and opportunity for improvement

S.4. **Numerator Statement:** Patients age 18 or older without preoperative carotid territory neurologic or retinal symptoms within the one year immediately preceding CEA who experience stroke or death during their hospitalization following carotid endarterectomy
S.7. **Denominator Statement:** Asymptomatic patients (based on NASCET criteria) on the within one year of CEA
S.10. **Denominator Exclusions:** DENOMINATOR EXCLUSIONS:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F OR Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Facility

Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

IF this measure is paired/grouped, NQF#/title:
De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Submitted SVS measure: Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Artery Stenting

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- The developer cites a variation in stroke or death following carotid endarterectomy, varying as a function of hospital volume in the Medicare population. Mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and 2.5% in low volume hospitals.

Updates:

The developer submitted additional evidence for this measure

- John J. Ricotta, MD, Ali AbuRahma, MD, FACS, Enrico Ascher, MD, Mark Eskandari, MD, Peter Faries, MD, Brajesh K. Lal, MD, Updated Society for Vascular Surgery guidelines for management of extracranial carotid disease, Society for Vascular Surgery. Journal of Vascular Surgery; September 2011,Volume 54, Issue 3, Pages e1–e31
- Additional references were submitted to support this measure.

Questions for the Committee:

If the developer provided updated evidence for this measure:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured*

Guidance from the Evidence Algorithm

Health outcome (Box 1) → relationship between health outcome and one action (Box 2) → Pass

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer cites a study where mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and 2.5% in low volume hospitals.
- Of 4,613 CEAs performed at 17 hospitals for asymptomatic patients in the VSGNE registry between 2003 to 2010, variation in postoperative stroke or death was 0% for the 25th quartile, 1.5% for the 75th quartile, with a median of 0.6%. The range across centers was 9% to 6.4%.
- Among 89 individual surgeons, postoperative stroke or death was 0% for the 25th quartile, 0.8% for the 75th quartile with a median of 0%. The range across surgeons was 0% to 25%.
- A 2016 analysis of the SVS VQI registry from 2010 to 2015 showed a median postoperative stroke or death rate of 0.4% compared to when the measure was developed where the median was 0.6%. The 75th percentile for centers was 1.8% which developers report as a slight increase.

Disparities

- 2016 analysis showed a slight variation in postoperative stroke or death rate by age between those younger than 60 (1.4%) and those older than 80 years (1.5%) versus individuals aged 60 - 79 (1.2% and 1.1%).
- The rate among females was 1.4%, compared to 1.1% for males.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is there expected variation in performance if reported at the physician level versus at the facility level?*

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

Committee pre-evaluation comments

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

1a.

- An appropriate outcome strongly linked to the index intervention.
- I would like to hear from the experts about the number of patients eligible for this measure -- it is very hard to make asymptomatic patients better. Is CEA still indicated?
- I note that the ASA has included a clone of this measure as a non-PQRS performance measure for anesthesiologists participating in the QCDR. Does SVS support this use? Should the associated anesthesia CPT code be listed in the denominator? Should the ASA be asked to contribute performance data as well, the next time this measure is assessed?
- The evidence providence applies directly to the measured outcome(s). Some clarification should be provided about what an asymptomatic carotid stenosis is since the statement as written after S.7 Denominator Statement: Asymptomatic patients (based on NASCET criteria) on the within one year of CEA-is difficult to understand as it is written. Also, treatment of asymptomatic carotid disease with CEA is typically based on the ACAS data since NASCET thresholds for citing surgical benefit was for symptomatic patients (we typically use the 'NASCET' method to measure the degree of ICA stenosis).
- -S.10 Denominator Exclusions: mentioned excluding patients w/ symptoms within 120 days- however the De. 3 Brief Description of Measure statement talks about the percentage of patients age > 18 or older w/o carotid territory neurologic or retinal symptoms within one year immediately preceding CEA…This seems inconsistent about what is being defined as asymptomatic. NASCET used 120 days and largely, patients w/o a hemispheric TIA/stroke/RIND event beyond 120 days are considered asymptomatic [usually not w/o symptoms for an entire year].
- However, most studies discuss 30-day stroke/death outcomes; the measure is focusing only on in-hospital stroke and death. There is some concern that only measuring in-hospital events will miss the proportion (1/3rd) of stroke/deaths that occur after hospital discharge but within 30-days (since LOS for most CEA cases is < 2 days). However, patients who develop in-hospital stroke/death have similar risk factors that predict out of hospital within 30-day stroke/death events so this seems to be a good quality indicator.

1b.

- This measure documents the most critical piece of information desired by patients contemplating this procedure; the measure is therefore appropriate for public accountability reporting even if variability is low.
- There is a moderate gap in performance demonstrated, with variability especially between low volume and high volume centers. Variability among surgeons in the registry is small, but median score is zero, so statistical power at the level of individual surgeons is likely very low.
- Performance data is provided using predominantly New England Vascular Study Group as well as SVS-VQI data. A philosophical question to consider is, clinically does a <1% variation between many of the various quartiles mentioned from the VQI data warrant a need for implementation for a quality measure? Also, the VQI records only about 10% of all the carotid procedures that occur in the US, the question remains as to whether the data from the VQI represent what is going on nationally.

1c.

- Measure is technically a construct, since it includes both stroke and mortality outcomes. Weighting these equally is reasonable, based on the seriousness of both outcomes. While in-house mortality or stroke is a reasonable measure of technical competence and therefore appropriate for evaluating surgeon (or anesthesiologist) performance, the numerator will slightly overstate the true rate of complications; patients with prolonged length of stay might suffer a stroke unrelated to surgical performance but prior to discharge. On the other hand, such a patient was probably a poor choice for surgery in the first place -- see my question about indications above -- and this is one aspect of performance which should be included in the measure.
- The Quality Construct seems reasonable however recording only in-hospital events will underestimate outcomes for stroke/death.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- **Data source(s):** Electronic Clinical Data: Registry. Registries used for this measure are The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE). VQI and VSGNE include hospitalization details and symptom status.

Specifications:

- This measure is specified at the clinician: group/practice and clinician: individual/facility level within the hospital/acute care facility setting.
- The numerator includes patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year of carotid endartectomy (CEA).
- The denominator includes patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year immediately preceding CEA.
- The developer notes they further specified the denominator by adding two category II performance reporting codes created since the last endorsement date.
- Patients are excluded if they are symptomatic or have other carotid stenosis less than 120 days prior to procedure.
- The measure is calculated as the number of asymptomatic patients undergoing CEA who have in-hospital stroke or death divided by the number of asymptomatic patients undergoing CEA.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate data elements and definitions included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Testing has not been updated since the previous endorsement, although the developer notes that the denominator has been updated. A summary of testing is provided below.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- A random sample of 100 patient records were reviewed, representing 5 relevant procedures from 5 different hospitals from data collected during the “past 2 years”. (Note that the ‘past 2 years’ refers to the time period prior to when the measure was first endorsed).
- Hospital reporting is proposed for every 12 months based on sufficient volume.
- Annual reporting of the last 50 consecutive procedures for surgeons (which may span more than one year) with suppression of <10 procedures is recommended.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the

VSGNE registry between 2003 to 2007.

- Chart abstraction was completed with results compared to registry data.
- Developers analyzed the level of agreement between the chart and registry data using the Kappa statistic.
- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched to patient data within the registry to compare discharge status (alive vs dead).
- Any discrepancies were further evaluated based on medical record audit.

Results of reliability testing

- Data element validity was used to support the reliability of the measure.
- Kappa statistics indicated strong agreement for identification of the correct procedure (CEA) performed (1.0), hospital mortality (.91), hospital stroke (1.0), and asymptomatic 120 days before treatment (.90).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) →Patient level data validity (Box 3) → (Box 10 of validity algorithm) →Appropriate method to assess data elements (Box 11) → Moderate certainty that data used in the measure are valid →Highest possible rating is moderate.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- Testing has not been updated since the previous endorsement. A summary of testing is provided below.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Method(s) of validity testing

- A random sample of 100 patient records were reviewed, representing 5 relevant procedures from 5 different hospitals.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the

VSGNE registry between 2003 to 2007.

Results of validity testing

- The developer reports that the percentage of asymptomatic patients treated in VSGNE (68%) and the postoperative stroke or death rate of 1.5% correspond to published data on asymptomatic patients.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients are excluded if they are symptomatic or have other carotid stenosis in less than 120 days prior to procedures
- Exclusions analysis on 862 asymptomatic patients undergoing elective CEA from the SVS registry, found a death rate of 0.7% and a stroke rate of 1.28% among 287 providers in 58 centers. The interquartile range for the combined endpoint was 0.2% to 7.6%.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is not risk adjusted.
- The developer gave the rationale that "risk adjustment is implicit within the quality measure" since the decision to perform this procedure "requires the interventionist to calculate the patient's risk-benefit ratio" .
- The developer provided a list of carotid endarterectomy definitions that includes a number of factors for pre-operative consideration.
- In addressing disparities, the developer states that such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region. The developer also reported slight variation in postoperative stroke or death rate by age between those younger than 60 (1.4%) and those older than 80 years (1.5%) versus individuals aged 60 - 79 (1.2% and 1.1%). The rate among females was 1.4%, compared to 1.1% for males.

Questions for the Committee:

- *Do you agree with the developer's rationale regarding risk adjustment?*
- *What is the Committee's expectation regarding consideration of SDS factors in maintenance measures?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Not provided.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- The developer reports no other data sources are available.

2b7. Missing Data

- The developer reports less than 1% missing data in VSGNE.

Guidance from the validity Algorithm
Precise specifications (Box 1) → All threats to validity assessed (Box 2) →Insufficient

Potential threats to vaidity around risk adjustment and SDS factors result in preliminary rating.

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

Committee pre-evaluation comments

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

2a.

- Both eligible procedures and measured outcomes seem well-specified. I would be interested to know how complication data are gathered in the two registries: Is this a harvest of ICD-10 codes from administrative data or a direct-eyeball assessment of the medical record? What are the criteria for documenting a stroke in the postop period, especially in a patient who might be comatose for multiple reasons?
- Case-mix adjustment is a difficult concept for asymptomatic carotid occlusive disease. The annualized stroke risk for a >70% ICA stenosis is <2% in the pre-statin era. The narrow therapeutic window of surgery vs. medical therapy for these lesions may lead some to argue that no risk adjustment should occur since it can 'wash away' bad decision making- e.g. if a patient has multiple comorbidities such as CHF, CRI, COPD and advanced age (>75), risk adjusting for the events may mask the concept that this patient should never have been offered carotid intervention for asymptomatic disease since their life-expectancy as well as the cost benefit analysis in patients w/ advanced age/CRI shows no benefit to surgery. This gets at the appropriateness of who should be offered prophylactic carotid revascularization.

2a2.

- Interesting to note that the kappa statistic is higher for in-house stroke than for in-house mortality. Why?
- Reliability testing appears adequate.

2b1.

- Specifications appear valid -- would be interested in more specificity on how stroke is defined and measured.
- The VSGNE cases were used for validity testing. This is a group of patients that are much different medically, socioeconomically and demographically from the rest of the patients in the VQI and likely nationally.

2b2.

- Strong face validity.
- Variation data presented seem consistent with expectations.
- What is the current overall penetration of registry participation among hospitals and vascular surgeons in the US?
- Validity testing was adequate in scope and number of patients.

2b3.

- No. This is core data for the registries where this measure will be reported.
- Threats to validity would be risk adjustment. In my opinion, prophylactic carotid revascularization should not be risk adjusted due to the very narrow therapeutic window of intervention and the known relatively benign natural history of the disease in the post-statin era.

2d.

- Yes and yes. I strongly agree with the decision NOT to risk adjust this measure.
- Stroke/death after asymptomatic carotid revascularization is a very important end-point and important quality indicator. I believe the analyses do demonstrate the aggregation and weighting rules for a quality construct and rationale.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data elements are generated and used by healthcare personnel during the provision of care, and coded by someone other than the person obtaining medical information
- The developer reports that VSGNE has tracked stroke or death since 2003 with percent missing for this variable at 1%.
- Developers report they have not experienced any difficulty obtaining this data from VSGNE.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are there fees to belong to the registry?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

- Requires individual assessment of medical records to ensure accurate diagnosis and capture of postop stroke.
- Death and stroke are reliably recorded for most in-patient stays. Many of these events occur outside of the hospitalization. Ideally, 30-day outcomes for both would be recorded however this is problematic since it may be difficult to link this to the antecedent hospitalization.

Criterion 4: Usability and Use

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

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

Current uses of the measure

- Centers for Medicare & Medicaid Services, Physician Quality Reporting System (measure #346)

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- PQRS is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). PQRS measures are used for public reporting on the Physician Compare website and for the quality component of the Value-Based Payment Modifier (VBPM).

Improvement results

- The developer reports that the number of centers and patients reported under this measure in VQI has grown over the last 5 years.
  - From the 5 year analysis of the VQI registry, 3.5% of the patients were from 2010 (rate of stroke or death, 1.5%) and 26.7% of the total were reported in 2015 (rate of stroke or death, 1.1%).
- Although the measure is reported in PQRS, these data were not provided by the developer

Unexpected findings (positive or negative) during implementation

- The developer states that data definitions regarding asymptomatic status based on NASCET criteria have eliminated confusion about symptoms.
- The developers also note that death is an accurate endpoint and that stroke has been accurately collected as judged by chart audits and comparison to claims data.

Potential harms

- The developer did not list any potential harms.

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

- This seems like a core measure for assessing the performance of surgeons and facilities. It has high face validity for public perception of this operation, and for patient choice of surgeon and hospital.
- PQRS measures are currently in place for asymptomatic carotid revascularization and are publically reported with current use of an accountability program.

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

- N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1540 NQF Project: Surgery Endorsement Maintenance 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

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

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)*:

discussed above

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The combined endpoint of stroke/death is the accepted primary endpoint for carotid endarterectomy. Variation in outcome has been established in randomized trials, cohort studies and meta analyses. This outcome measure has face validity among all providers of this service. Studies cited above have shown substantial variation in outcomes by provider when CEA is performed in asymptomatic patients.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*:

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Expert opinion.

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Stroke/death after CAS is the reporting standard recommended by the Society for Vascular Surgery, and has been used in multiple RCTs.

1c.14 Summary of Controversy/Contradictory Evidence: None

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1. Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Jama 1995;273(18):1421-8.

2. Halliday A, Mansfield A, Marro J, et al. Prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms: randomised controlled trial. Lancet 2004;363(9420):1491-502.

3. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med 1991; 325: 445–53.

4. Biller J, Feinberg WM, Castaldo JE, et al. Guidelines for carotid endarterectomy: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. Stroke; a journal of cerebral circulation 1998;29(2):554-62.

5. Kresowik TF, Bratzler DW, Kresowik RA, et al. Multistate improvement in process and outcomes of carotid endarterectomy. J Vasc Surg 2004;39(2):372-80.

6. Wennberg DE, Lucas FL, Birkmeyer JD, Bredenberg CE, Fisher ES. Variation in carotid endarterectomy mortality in the Medicare population: trial hospitals, volume, and patient characteristics. Jama 1998;279(16):1278-81.

7. Feasby TE, Kennedy J, Quan H, Girard L, Ghali WA. Real-world replication of randomized controlled trial results for carotid endarterectomy. Archives of neurology 2007;64(10):1496-500.

8. Cronenwett JL, Likosky DS, Russell MT, Eldrup-Jorgensen J, Stanley AC, Nolan BW. A regional registry for quality

assurance and improvement: The Vascular Study Group of Northern New England (VSGNNE). J Vasc Surg 2007.

9. Tu J, Wang H, Bowyer B, Green L, Fang J, Kucey D. Risk Factors for Death or Stroke After Carotid Endarterectomy: Observations From the Ontario Carotid Endarterectomy Registry. Stroke. 2003;34:2568-2575.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:

Biller J, Feinberg WM, Castaldo JE, et al. Guidelines for carotid endarterectomy: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. Stroke; a journal of cerebral circulation 1998;29(2):554-62.

2016 Addition from Society for Vascular Surgery September 2011 Guidelines

Management of carotid bifurcation stenosis is a cornerstone of stroke prevention and has been the subject of extensive clinical investigation, including multiple controlled randomized trials. The appropriate treatment of patients with carotid bifurcation disease is of major interest to the community of vascular surgeons. In 2008, the Society for Vascular Surgery published guidelines for treatment of carotid artery disease. At the time, only one randomized trial, comparing carotid endarterectomy (CEA) and carotid stenting (CAS), had been published. Since that publication, four major randomized trials comparing CEA and CAS have been published, and the role of medical management has been re-emphasized.

1c.17 Clinical Practice Guideline Citation: Biller J, Feinberg WM, Castaldo JE, et al. Guidelines for carotid endarterectomy: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. Stroke; a journal of cerebral circulation 1998;29(2):554-62.

2016 Addition:

John J. Ricotta, **MD, Ali AbuRahma, MD, FACS, Enrico Ascher, MD, Mark Eskandari, MD, Peter Faries, MD, Brajesh K. Lal, MD, Updated Society for Vascular Surgery guidelines for management of extracranial carotid disease, Society for Vascular Surgery. Journal of Vascular Surgery;** September 2011**Volume 54, Issue 3, Pages e1–e31**

1c.18 National Guideline Clearinghouse or other URL: N/A

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: AHA

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: Level 1

1c.24 Rationale for Using this Guideline Over Others: Universally accepted

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: 1c.26 Quality: 1c.27 Consistency:

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1540_Evidence_MSF5.0_Data_2016.doc

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Numerous manuscripts have noted variation in the combined endpoint of stroke or death following carotid endarterectomy. In the Medicare population, the outcome has been shown to vary substantially as a function of hospital volume. This is an important consideration, since it is widely recognized that many surgeons and centers performing CEAs do not meet the high standards of the randomized trials which established the benefit of such treatment. It has been shown that mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and fully 2.5% in low volume hospitals (Ref 6). Given that the stroke rate is generally 3 times the mortality rate, this suggests that some centers/surgeons are not achieving optimal results. A recent survey in Canada found that 45% of hospitals are not meeting published guidelines (Ref 7). Adoption of this outcome measure in the United States would likely disclose similar results and lead to quality improvement when this information was provided to surgeons and centers. This effect has been demonstrated in a midwest regional study by Kresowik et al where stroke and death rate after CEA improved after providing outcome data (Ref 5). The VSGNNE has shown that regional results are good for CEA outcomes, but significant variation does exist between surgeons and centers (Ref 8). Postoperative stroke or death is the accepted outcome paramenter for this surgery, and its measurement and reporting would demonstrate variation and opportunity for improvement

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
It has been shown that mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and fully 2.5% in low volume hospitals (Ref 6). Given that the stroke rate is generally 3 times the mortality rate, this means that many ill advised operations are likely being performed. A recent survey in Canada found that 45% of hospitals are not meeting published guidelines (Ref 7).

For this measure propsal we reviewed 4,613 CEAs performed for asymptomatic patients in VSGNE between 2003 to 2010. Among 17 hospitals, the variation in postoperative stroke or death rate was as follows: The 25th quartile was 0%. The 75th quartile was 1.5%. The median was 0.6%. The range across centers was 0% to 6.4%. Similarly, among 89 individual surgeons the rates were as follows: The 25th quartile was 0%. The 75th quartile was 0.8%. The median was 0%. The range across surgeons was 0% to 25%. This demonstrates substantial variability and performance gap even though the regional average outcome was excellent.

In 2016, an analysis was run on the data for this measure collected in the SVS VQI registry from 2010 - 2015. This analysis based on 261 centers representing 1,251 physicians and 27,773 cases demonstrated a slightly better median at 0.4% versus the median when this measure was created of 0.6%. However the 75% percentile for centers was 1.8% in the five years of data since this measure was created, demonstrating a slight increase. Therefore, we continue to see variation as was noted above when the measure was created.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
See list in 1a.4 above

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region.

The data from our 2016 analysis showed very little variation in this measure in regard to disparities. There was a very slight variation

by age between those younger than 60 (1.4%) and those older than 80 year old (1.5%) versus those individuals between the ages of 60 - 79 (1.2% and 1.1%). There was no difference regarding race and again a very slight difference regarding gender with females at 1.4% and males at 1.1%.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
not available

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Stroke or death following CEA has been the primary clinical endpoint for multiple randomized trials of CEA (Ref 1-3). Although this is sometimes reported after 30 days, most postoperative strokes or deaths occur in hospital following CEA for asymptomatic patients (Ref 1). This endpoint is easy to capture from claims data and registries. This outcome is particularly important for asymptomatic patients undergoing CEA, since this is a prophylactic operation being proposed to prevent future stroke. As such, guidelines from the American Heart Association recommend CEA for such patients only if the risk of surgical death or stroke combined is less than 3% (Ref 4). This is based on Level I evidence from randomized trials which established the benefit of CEA in asymptomatic patients with at least 60% internal carotid artery (ICA) stenosis, but only if the surgical risk is appropriately low, since the subsequent stroke risk with medical management is not high (Ref 1-2). This contrasts with symptomatic patients with severe ICA stenosis where the stroke risk under medical therapy is high, and justifies CEA even when stroke risks are higher.

Stroke is defined as an acute neurological deficit due to an occlusive or hemorrhagic brain lesion that persists more than 24 hours. It can be substantiated by a new stroke seen on brain imaging, but this is not a requirement, i.e., clinical symptoms alone is sufficient. Both minor and major strokes will be counted, as long as the symptoms persist more than 24 hours. Stroke in either carotid distribution, or vertebrobasilar stroke is included, i.e., any postoperative new neurologic deficit attributed to an occlusive or hemorrhagic brain lestion lasting more than 24 hours. From an operational standpoint, post-operative new stroke is defined by medical record coding, ICD-9-CM 997.02.

While stroke or death following CEA is an appropriate quality measure for either symptomatic or asymptomatic patients, we believe that the former group would require risk adjustment to allow fair comparisons, while we do not believe this is necessary for asymptomatic patients. The rationale for this is as follows. Factors such as atrial fibrillation, congestive heart failure, contralateral carotid occlusion and diabetes have been shown to increase stroke risk following CEA, in addition to symptom status, and could be used to justify risk stratification (Ref 9). However, for asymptomatic patients, it is incumbent upon the surgeon to select only those patients of low perioperative risk to benefit from CEA. In fact, the recommendations of the AHA are that this surgery should not be done if risk is high (>3%), without risk adjustment in asymptomatic patients (Ref 4).

We propose that the denominator for this measure should be patients who have never been symptomatic in either the cerebral hemisphere ipsilateral to the carotid lesion, the contralateral hemisphere or the vertebrobasilar circulation(dizziness or lightheadedness alone are not considered symptoms). This group has the lowest risk of stroke with carotid intervention and also the lowest risk of stroke with medical therapy alone.

Adopting this outcome measure would likely have immediate impact on improving quality. Regional data have shown that feedback of the key outcome of stroke and death, in addition to some process measures after CEA reduced this outcome from 5.6% to 5.0% and in asymptomatic patients from 4.1% to 3.8% (Ref 5). The reporting time frame for hospitals should be on a yearly basis. The time frame for surgeons should be cumulative over their career.

This is an important quality measure, since it is suspected that a number of surgeons and centers performing CEAs do not meet the high standards of the randomized trials which established the benefit of such treatment. It has been shown that mortality following CEA in Medicare patients was 1.4% in hospitals participating in randomized trials, 1.7% in high volume non-trial hospitals, 1.9% in average volume hospitals and fully 2.5% in low volume hospitals (Ref 5). Given that the stroke rate is generally 3 times the mortality rate, this means that some surgeons/centers are likely not achieving optimal results. A recent survey in Canada found that 45% of hospitals are not meeting published guidelines (Ref 7). Adoption of this outcome measure in the United States would likely disclose similar results and lead to quality improvement. The VSGNNE has shown that regional results are good for CEA outcomes, but significant variation does exist between surgeons and centers (Ref 8). This would be the first true outcome measure for vascular surgery, and it would apply to the most frequently performed vascular operation.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1. Endarterectomy for asymptomatic carotid artery stenosis. Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Jama 1995;273(18):1421-8.
2. Halliday A, Mansfield A, Marro J, et al. Prevention of disabling and fatal strokes by successful carotid endarterectomy in patients without recent neurological symptoms: randomised controlled trial. Lancet 2004;363(9420):1491-502.
3. North American Symptomatic Carotid Endarterectomy Trial Collaborators. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med 1991; 325: 445–53.
4. Biller J, Feinberg WM, Castaldo JE, et al. Guidelines for carotid endarterectomy: a statement for healthcare professionals from a special writing group of the Stroke Council, American Heart Association. Stroke; a journal of cerebral circulation 1998;29(2):554-62.
5. Kresowik TF, Bratzler DW, Kresowik RA, et al. Multistate improvement in process and outcomes of carotid endarterectomy. J Vasc Surg 2004;39(2):372-80.
6. Wennberg DE, Lucas FL, Birkmeyer JD, Bredenberg CE, Fisher ES. Variation in carotid endarterectomy mortality in the Medicare population: trial hospitals, volume, and patient characteristics. Jama 1998;279(16):1278-81.
7. Feasby TE, Kennedy J, Quan H, Girard L, Ghali WA. Real-world replication of randomized controlled trial results for carotid endarterectomy. Archives of neurology 2007;64(10):1496-500.
8. Cronenwett JL, Likosky DS, Russell MT, Eldrup-Jorgensen J, Stanley AC, Nolan BW. A regional registry for quality assurance and improvement: The Vascular Study Group of Northern New England (VSGNNE). J Vasc Surg 2007.
9. Tu J, Wang H, Bowyer B, Green L, Fang J, Kucey D. Risk Factors for Death or Stroke After Carotid Endarterectomy: Observations From the Ontario Carotid Endarterectomy Registry. Stroke. 2003;34:2568-2575.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.vascularqualityinitiative.org/wp-content/uploads/2016_PQRS_Information-v2-1.pdf

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** CEA defs v.01.09.doc

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
We have increased the specificity of the denominator by having two category II performance reporting codes created since the last endorsement date.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients age 18 or older without preoperative carotid territory neurologic or retinal symptoms within the one year immediately preceding CEA who experience stroke or death during their hospitalization following carotid endarterectomy

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Since hospitals have sufficient annual volume to generate accurate reporting levels, these are proposed for reporting every 12 months for hospital. Since surgeons have lower individual volume, we recommend annual reporting of the last 50 consecutive procedures, which may span more than one year, with suppression if < 10 procedures (ie, reported as too low volume to report).

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
ANY registry that includes hospitalization details and symptom status within 120 days is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. If a registry collects this data then they could report this measure. Patients who were asymptomatic within one year of the CEA (CPT code 37215) who died or experienced postoperative in hospital stroke are included.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Asymptomatic patients (based on NASCET criteria) on the within one year of CEA

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
ANY registry that includes hospitalization details and symptom status within 120 days is required to identify patients for denominator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Patients who were asymptomatic within one year of the CAS (CPT code 37215)are included.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
DENOMINATOR EXCLUSIONS:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F
OR

Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
DENOMINATOR EXCLUSIONS:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F
OR
Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not required

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See "Scientific Acceptablility" section for rationale

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Asymptomatic patients undergoing CEA who experience inhospital stroke or death/all asymptomatic patients undergoing CEA.

This measure is to be reported each time a CEA is performed during the reporting period. It is anticipated that clinicians who provide services of CEA, as described in the measure, based on the services provided and the measure-specific denominator coding will report this measure. This measure may be reported by clinicians who perform the quality actions described in the measure based on the services provided and the measure-specific denominator coding.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample*

size.)
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Society for Vascular Surgery Vascular Quality Initiative Registry
Vascular Study Group of New England Registry

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1540_MeasureTesting_MSF5.0_Data_v1.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1540 NQF Project: Surgery Endorsement Maintenance 2010

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)

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 field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

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)*:
A random sample of 100 patient records representing 5 procedures relevant to the measure from 5 different hospitals based on data collected during the past 2 years. In addition, in-hospital mortality was examined by claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007. These measures were originally tested in 2011 and this was the most recent data. All of the testing was approved by the NQF Steering Committee at the time that the measures were first approved in 2012. These measures are approved for PQRS reporting and working well. Regarding the sample and the data, this is an accepted testing practice to pull a sample for chart review to then compare to the data that was submitted to a registry.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
A nurse abstractor completed a form based on medical record review for the variables relevant to this measure. The results of this chart review were then compared with the original registry data. The Kappa statistic was used to judge reliability of the data. For mortality validation, claims data from each of 12 hospitals were matched to patient identified data within the VSGNE registry to compare discharge status (alive vs. dead). Any discrepencies were then further evaluated based on a medical record audit.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The key variables for this measure and testing results were:

1. Correct procedure (carotid endarterectomy) performed. Kappa =1.0
2. Hospital mortality: Kappa = .91 (SE .01)
3. Hospital stroke: Kappa = 1.0
4. Asymptomatic 120 days pre-Rx: Kappa = .90 (SE .07)

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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 reliability testing

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
Comparison of results with expected results from literature. Please see the evidence listed in the NQF form under importance.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
The percentage of asymptomatic patients being treated with CEA in VSGNE of 68% corresponds to published data on this cohort. The postop stroke or death rate of 1.5% also correponds to published results for asymptomatic patients.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(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)*:
SVS Vascular Registry 862 asymptomatic patients undergoing elective CEA

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient*

preference):
measure calculation

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
Death rate 0.7%, stroke rate 1.28% among 287 provider in 58 centers
Interquartile range was 0.2-7.6% for the combined endpoint

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 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 "Scientific Acceptablility" section for rationale. Risk adjustment is implicit within this quality measure as judged by the sponsor, the Society for Vascular Surgery, for the following reason. CEA in an asymptomatic patients is a prophylactic procedure designed to prevent future stroke. The decision to perform such a procedure requires the interventionist to calculate the patient´s risk-benefit ratio, in order to avoid post-CEA stroke or death that eliminate the benefit of the procedure. Risk adjustment based on patient factors should not be applied, since high risk patients should not undergo this prophylactic procedure, and using risk adjustment would reward interventionists who selected high risk patients for treatment.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
see section 1.b.3 and above 2,d,5

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
Standard statistial analysis to determine 95% confidence interval for hospitals and providers to determine practical difference from mean

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
other sample not available

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: N/A

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain: Disparities have not been reported. Please see the new data under the importance sections of the NQF regular form.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met? (*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition,
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
Yes

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
In the VSGNE experience which has been tracking stroke or death as a major endpoint since 2003, we have not experienced any difficulty with obtaining data related to this endpoint. Our percent missing for this variable has been less than 1%.

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Payment Program PQRS www.cms.hhs.gov

4a.1. **For each CURRENT use, checked above, provide:**

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

This measure has been accepted into the PQRS reporting program as PQRS measure number 346. Its purpose is for quality reporting by physicians into Medicare and it is a national program.

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

The number of centers and patients reported under this measure has grown over the last 5 years as VQI has grown. In our 5 year analysis, 3.5% of the patients were from 2010 while 26.7% of the total were reported in 2015. In 2010 the rate of stroke or death was 1.5% while in 2015 it was 1.1%.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
Data definitions regarding asymptomatic status based on NASCET criteria have eliminated confusion about symtoms. Death is an accurate endpoint. Stroke has been accurately collected as judged by chart audits and comparison to claims data that has been done within VSGNE.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Society for Vascular Surgery
Co.2 **Point of Contact:** Sarah, Murphy, ladams@vascularsociety.org, 312-334-1229-
Co.3 **Measure Developer if different from Measure Steward:** Society for Vascular Surgery
Co.4 **Point of Contact:** Jill, Rathbun, Jill_Rathbun@galileogrp.com, 312-334-2305-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

CAROTID ENDARTERECTOMY DEFINITIONS – v.01.09

If more than one response applies, select the most severe (highest number) response for each data field.

Pre-op

Smoking: Prior = quit ≥ 1 year ago. Current = still smoking within last 12 months. Include cigarettes, pipe, or cigar.

HTN (Hypertension): Defined as ≥ 140/90, either systolic or diastolic, at admission or within last 6 months, or clearly documented in medical record.

Beta-blockers: Peri-operative = started within one month before surgery or during surgery. Chronic = more than one month before surgery.

CAD Symptoms (Coronary artery disease**):** Stable angina = stable pattern or symptoms with or without anti-anginal medication. Unstable angina = new onset, increasing frequency, lasting > 20 min and/or rest angina.

CABG/PTCA: Coronary artery bypass, angioplasty, or stent.

CHF (Congestive Heart Failure): Documented CHF: Mild = SOB on exertion; Severe = SOB at rest, pulmonary edema, or pitting ankle edema. (Use 2 = mild if severity not documented.)

COPD: Not treated = COPD documented in record but not treated with medication. Medication includes theophylline, aminophylline, inhalers or steroids

Dialysis: Transplant = patient has functioning kidney transplant; Dialysis = currently on hemo- or peritoneal dialysis.

Creatinine: Last available measurement taken before procedure. If multiple measurements, use highest within 30 days of surgery.

Stress Test: Includes stress EKG, stress echo, nuclear stress scans, within 2 years of surgery.

Previous Arterial:

Bypass - Any non-cardiac arterial bypass for occlusive disease
CEA - Carotid endarterectomy
Aneurysm Repair – Any known true arterial aneurysm repair (excluding cerebral or pseudo-aneurysm)
PTA/Stent – Of any non-cardiac artery
Major Amputation – Any amputation above the foot or hand

Pre-admin living: Use last living status before any current, acute hospitalization or rehab unit.

Pre-Op Medications: Taken within 36 hours of surgery. Statins include any HMG-CoA reductase inhibitor, such as Lipitor, Mevacor, Pravachol, Zocor, Lescol, etc. If Plavix is discontinued prior to surgery it should be coded = 0.

Pre-op Hemoglobin: Most recent pre-op hemoglobin within past 30 days.

Symptoms: Ocular: unilateral visual loss or major blurring, etc. Cortical: unilateral motor and/or memory loss, or dysphagia/aphasia, etc. Vertebrobasiliar: bilateral motor, sensory, or visual loss, diplopia, ataxaia, etc. Major cortical or vertebrobasilar stroke = disability causing non-independent living status. Minor stroke is non-disabling. Major ocular stroke = blindness, otherwise minor. Stroke<1 month means stroke within previous month before surgery, etc. TIA=transient ischemic attack completely resolved within 24 hours.

Non-specific: Not clearly a carotid or vertebrobasilar TIA, e.g., light-headedness, dizziness

Ipsilat stroke on CT/MRI: Carotid territory only.

CEA: Carotid endarterectomy

Previous radiation: Radiation therapy in a field including the affected carotid artery.

ICA stenosis: Use most severe category by modality thought to be most accurate if multiple modalities used.

Procedure

Urgency: Urgent = surgery within 24 hrs of admit or patient can't be discharged; emergent = surgery within 6 hrs of admission.

Shunt: If used, specify if routinely used (1), or if placed selectively in this patient for a specific indication (2).

Re-explore artery after closure: for defect detected after closure during same operation.

Concomitant Procedure

Proximal endovascular: Angioplasty or stent of more proximal carotid, innominate artery.

Post-op

Cranial nerve injury: Any occurrence, transient or persisting: VII-facial droop or more severe; IX-swallowing difficulty unless other diagnosis confirmed; X- hoarseness unless laryngoscopy normal; XII-any tongue deviation or dis-coordination

Ipsilat/Contralat neurologic event: Cerebral or ocular. TIA = cortical or ocular symptoms <24hrs duration. Major cortical or vertebrobasilar stroke = disability causing non-independent living status. Otherwise, minor. Major ocular stroke = blindness, otherwise minor. Minor stroke is non-disabling.

Time of Onset Ipsila/Contralat: Time when first noticed, but if noted on awakening from anesthesia code as 1=intra-op. Use 2=≤6 hrs post-op if normal at completion of procedure, and then neurologic event developed.

Reperfusion Symptoms: Seizures associated with headache, or hemorrhage on CT/MRI.

IV meds required: Indicates continuous infusion or more than one dose required more than one hour after surgery.

Myocardial Infarction: Troponin: by local standards for MI. EKG: new Q waves, new ST and T wave changes. Clinical: documentation of MI by clinical criteria or ECHO or other imaging modality.

Dysrhythmia: New rhythm disturbance requiring treatment with medications or cardioversion.

CHF: Pulmonary edema with requirement for monitoring or treatment in ICU.

Return to OR for bleeding: Applies to carotid endarterectomy incision only. Use 666 if Return to OR = 0.

Return to OR for Neurologic Event: Use 666 if Return to OR = 0.

Peri-operative Antibiotics: Use 0=no if antibiotic was not ordered. To use 1=yes, antibiotic must be ordered to be given within 1 hour prior to skin incision and must be ordered to be discontinued within 24 hrs of end of time of operation. To use 2=no for medical reason, a medical reason must be documented in the chart that antibiotic not given. **Acceptable antibiotics include:** Ampicilin/sulbactam, Aztreonam, Cefazolin, Cefmetazole, Cefotetan, Cefuroxime, Ciprofloxacin, Clindamycin, Ertapenem, Erythromycin base, Gatifloxacin, Gentamicin, Levofloxacin, Metronidazole, Moxifloxacin, Neomycin, and Vancomycin.

1st-2nd Generation Cepahalosporin: (Cefazolin or Cefuroxime) Use response 1=yes, if ordered. If documented in medical record that not ordered for medical reason use 2. Otherwise use 0=no.

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1543
De.2. **Measure Title:** Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Artery Stenting (CAS)
Co.1.1. **Measure Steward:** Society for Vascular Surgery
De.3. **Brief Description of Measure:** Percentage of patients 18 years of age or older without carotid territory neurologic or retinal symptoms within 120 days immediately proceeding carotid angioplasty and stent (CAS) placement who experience stroke or death during their hospitalization for this procedure. This measure is proposed for both hospitals and individual interventionalists.
1b.1. **Developer Rationale:** Better patient selection to avoid treating high risk patients who will likely experience stroke or death after CAS for asymptomatic patients which eliminates any benefit of the procedure.

S.4. **Numerator Statement:** Patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year of their procedure who experience stroke or death during their hospitalization following elective carotid artery angioplasty and stent placement.
S.7. **Denominator Statement:** Patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year immediately preceding carotid artery stenting.
S.10. **Denominator Exclusions:** Per PQRS Specifications for 2016: DENOMINATOR EXCLUSIONS:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F
OR Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Clinician : Individual, Facility

Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

IF this measure is paired/grouped, NQF#/title:
De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** Submitted SVS measure: Postoperative Stroke or Death in Asymptomatic Patients undergoing Carotid Endarterectomy

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported

by the stated rationale.

Summary of evidence:

- The developer notes better patient selection for CAS is needed for asymptomatic patients to avoid treating patients at higher risk for stroke or death following CAS.
- Guidelines from the American Heart Association recommend carotid endartectomy (CEA) for patients if the risk of death or stroke is less than 3%; the developers state this same threshold for asymptomatic patients undergoing CAS although could also be used although there is no published guideline for CAS.
- Cochrane Database analysis of stroke or death within 30 days of CAS for asymptomatic carotid stenosis showed no difference between CEA and CAS in all patients as well as for a subset of patients deemed not suitable for surgery.
- CAPTURE-2 and EXACT stent trials demonstrated outcomes for CAS in asymptomatic patients that were comparable to those established by the AHA for patients treated with CEA.
- Additional citations for evidence are linked here.

Updates

The developer submitted updated evidence for this measure

- "Safety of stenting and endarterectomy by symptomatic status in the Carotid Revascularization Endarterectomy Versus Stenting Trial (CREST)." Silver FL(1), Mackey A, Clark WM, Brooks W, Timaran CH, Chiu D, Goldstein LB, Meschia JF, Ferguson RD, Moore WS, Howard G, Brott TG; CREST Investigators. Stroke. 2011 Mar;42(3):675-80. doi: 10.1161/STROKEAHA.110.610212.
- "Randomized Trial of Stent versus Surgery for Asymptomatic Carotid Stenosis". Rosenfield K(1), Matsumura JS(1), Chaturvedi S(1), Riles T(1), Ansel GM(1), Metzger DC(1), Wechsler L(1), Jaff MR(1), Gray W(1); ACT I Investigators. N Engl J Med. 2016 Mar 17;374(11):1011-20. doi: 10.1056/NEJMoa1515706. Epub 2016 Feb 17.
- "Experience matters more than specialty for carotid stenting outcomes" Sgroi, Michael D. et al. Journal of Vascular Surgery 2015, Volume 61, Issue 4, 933 - 938.
- Experience and outcomes with carotid artery stenting: an analysis of the CHOICE study (Carotid Stenting for High Surgical-Risk Patients; Evaluating Outcomes Through the Collection of Clinical Evidence).JACC Cardiovasc Interv. 2014 Nov;7(11):1307-17. doi: 10.1016/j.jcin.2014.05.027.

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Guidance from the Evidence Algorithm
Health outcome (Box 1)→relationship between outcome and one action is supported by rationale (Box 2) → Pass
Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

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

- Data by the developer in the previous submission show that CAS procedures were completed for asymptomatic patients in 65% of patients in VSGNE undergoing CAS.
- An analysis of 2010-2015 VQI self-reported data (175 centers, 544 providers, 3,342 procedures) showed stroke or death within 30 days after CAS to be 2.1%.
    - The developer also notes a decrease in the percentage of cases with a reported death within 30 days of CAS has fallen from 2% to 1.6%, with the exception of in 2012 where the rate was 3.3% .
    - By center over 2010-2015, the interquartile range was 0% to 1.7% per center with the number of centers increasing each year.

Disparities

- In the 2016 analysis of 3,342 patients reported from 2010-2015 in the SVS VQI, the developer reports that the patients experiencing stroke or death within 30 days of a CAS procedure were older, had Medicare, and were slightly more likely to be female. The developer notes they did not see specific differences by race.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Is there expected variation in performance if reported at the physician level versus at the facility level?*

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

Committee pre-evaluation comments

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

1a. Evidence to Support Measure Focus

- An appropriate outcome strongly linked to the index intervention.

  I would like to hear from the experts about the number of patients eligible for this measure -- it is very hard to make asymptomatic patients better. Is CAS really indicated over statin therapy/observation?

  Are there data available on the risk for stroke/death during observation based on patient demographics, size or location of the plaque? It seems these should be part of the indications for the procedure, otherwise every healthy patient over 18 would be eligible!

  As with many advanced medical procedures, the best results are likely to arise in populations that don't need the intervention -- these groups have fewer co-morbidities and better outcomes with or without the procedure. How do the developers intend to address the risk of overuse?
- The evidence for the outcome applies directly. The proposed outcome of stroke/death are standard and important quality indicators of carotid intervention. The relationship between the measured outcome (in-hospital stroke/death) are supported by the stated rationale since appropriate patient selection is crucial to getting the stated benefit of the procedure.

1b. Performance Gap

- This measure documents the most critical piece of information desired by patients contemplating this procedure; the measure is therefore appropriate for public accountability reporting even if variability is low. 2) There is a moderate gap in performance demonstrated, with variability especially between low volume and high volume centers. Variability among surgeons in the registry is small, but median score is zero, so statistical power at the level of individual surgeons is likely very low.

  Variability data presented are for 30-day stroke risk, but the measure is specified to hospital discharge, which might be only 1 day. Can the developers shed light on the likely difference between 30 day and 1 day outcomes?

- The volume-outcome relationship and provider experience with CAS is well-known. It will be important to consider these features if reporting of the measure is endorsed. CAS generally carries a 2-fold greater risk of stroke/death even for asymptomatic patients when compared to CEA in a non-trial environment with variably experienced providers. The tremendous variability in patient selection and outcome for CAS even for asymptomatic disease represents an important performance gap. It is not clear that the VQI sample of CAS procedures reflects national patterns in utilization and outcomes.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- **Data source(s):** Electronic Clinical Data: Registry. Registries used for this measure are The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE). VQI and VSGNE include hospitalization details and symptom status.

Specifications:

- This measure is specified at the clinician: group/practice, clinician: individual/facility level for the hospital/acute care facility setting.
- The numerator includes patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year of their procedure who experience stroke or death during their hospitalization following elective carotid artery angioplasty and stent placement.
- The denominator includes patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year immediately preceding carotid artery stenting.
- Patients are excluded if they are symptomatic or have other carotid stenosis less than 120 days prior to procedure.
- The measure is calculated as the number of asymptomatic patients undergoing CAS who have in hospital stroke or death divided by the number of asymptomatic patients undergoing CAS.
- The developer notes that this measure is to be reported each time a CAS is performed during the reporting period.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate data elements and definitions included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Testing has not been updated since the previous endorsement. A summary of testing is provided below.

SUMMARY OF TESTING

Reliability testing level ☐ **Measure score** ☒ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- A random sample of 100 patient records was reviewed, representing 5 relevant procedures from 5 different hospitals from data collected during the “past 2 years”. (Note that the ‘past 2 years’ refers to the time period prior to when the measure was first endorsed).
- Hospital reporting is proposed for every 12 months based on sufficient volume.
- Annual reporting of the last 50 consecutive procedures for surgeons (which may span more than one year) with suppression of <10 procedures is recommended.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007.
- Chart abstraction was completed with results compared to registry data.
- Developers analyzed the level of agreement between the chart and registry data using the Kappa statistic.

- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched to patient data within the registry to compare discharge status (alive vs dead).
- Any discrepancies were further evaluated based on medical record audit.

Results of reliability testing

- Data element validity was used to support the reliability of the measure.
- Kappa statistics indicated strong agreement for identification of the correct procedure (CAS) performed (1.0), hospital mortality (.91), hospital stroke (1.0), and asymptomatic 120 days before treatment (.90).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- *Is the data element level testing provided enough to also confirm reliability and validity for physician/clinician level performance?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → Empirical reliability testing (Box 2) →Patient level data validity (Box 3) → (Box 10 of validity algorithm) →Appropriate method to assess data elements (Box 11) → Moderate certainty that the data used in the measure are valid →Highest possible rating is moderate.

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- Testing has not been updated since the previous endorsement. A summary of testing is provided below.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☒ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Method(s) of validity testing

- A random sample of 100 patient records was reviewed, representing 5 relevant procedures from 5 different hospitals.
- In-hospital mortality was examined using claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007.
- Chart abstraction was completed with results compared to registry data.
- For mortality validation, claims data from each of 12 hospitals participating in the VSGNE registry were matched to patient data within the registry to compare discharge status (alive vs dead).

Results of validity testing

- Kappa statistics indicated strong agreement for identification of the correct procedure (CAS) performed (1.0), hospital mortality (.91), hospital stroke (1.0), and asymptomatic 120 days before treatment (.90).
- The developer reports that the percentage of asymptomatic patients treated in VSGNE (60%) and the postoperative stroke or death rate of 2.2% correspond to published data on asymptomatic patients.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- Patients are excluded from the measure if they have NASCET criteria symptoms within the one year proceeding CAS; are symptomatic carotid stenosis less than 120 days prior to procedure, or other carotid stenosis 120 days or greater prior to procedure.
- Exclusions analysis was completed on 805 asymptomatic patients undergoing elective CEA from the SVS registry.
  - Death rate of 2.0% and a stroke rate of 2.11% among 287 providers in 58 centers. The interquartile range for the combined endpoint was 0.3% to 8.6%.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

Risk adjustment summary

- The measure is not risk adjusted.
- The developer gave the rationale that "risk adjustment is implicit within this quality measure" since the decision to perform this procedure "requires the interventionist to calculate the patient's risk-benefit ratio".
- The developer provided a list of carotid artery stenting definitions that includes a number of factors for pre-operative consideration.
- In addressing disparities, the developer states that such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region. In analysis of disparities, the developer reports that the patients experiencing stroke or death within 30 days of a CAS procedure were older, had Medicare, and were slightly more likely to be female. The developer notes they did not see specific differences by race.

Questions for the Committee:

- Do you agree with the developer's rationale regarding risk adjustment?
- What is the Committee's expectation regarding consideration of SDS factors in maintenance measures?

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Data by the developer in the previous submission show that CAS procedures were completed for asymptomatic patients in 65% of patients in VSGNE undergoing CAS.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- The developer reports no other data sources are available.

2b7. Missing Data

- The developer notes less than 1% missing data from both VSGNE and VQI.

Guidance from the Validity Algorithm
Precise specifications (Box 1) → All threats to validity assessed (Box 2) →Insufficient

Potential threats to validity around risk adjustment and SDS factors result in preliminary rating.

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications: Reliability-Specifications

- I'd be interested in a little more information on this statement:
  Patients are excluded from the measure if they have NASCET criteria symptoms within the one year proceeding CAS; are symptomatic carotid stenosis less than 120 days prior to procedure, or other carotid stenosis 120 days or greater prior to procedure.

  Is it asymptomatic for a year before CAS, or 120 days?

- No issues with the data elements other than proposing the question as to whether or not the number of CAS procedures in the VQI truly reflects national trends in utilization and outcomes. Would favor non-risk adjusted data given relatively benign natural history of asymptomatic carotid disease with the narrow therapeutic window for carotid revascularization procedures. The reliability algorithm is clearly presented and the data element validity testing has moderate to high acceptability.

2a2. Reliability - Testing:

- Again, kappa is greater for stroke definition than for mortality. What's up with that? Both scores are high, however, lending support to the reliability of data capture, at least in the measured population.

- Sample size was modest- unclear if the VQI CAS patients reflect national trends in utilization and outcomes. Moderate reliability was the preliminary rating assigned.

2b.1 Validity – Specifications

- Highly valid measure from the public perspective and also from the clinician's point of view. This measure has very good alignment between clinical and patient-centered goals.\

- No specific threats are identified.

2b2. Validity - Testing

- Again, what is the participation in the registry of facilities and providers doing CAS?

  Assuming decent and increasing penetration, then this measure has high face validity -- easy to understand and highly relevant.

- Validity testing was adequate (100 random sample patients, 7205 VSGNE patients checked for mortality; direct chart abstraction w/ kappa statistic testing was excellent between procedure identification and mortality/stroke outcomes in the claims data and VSGNE data.

2b3-7. Threats to Validity

- The greatest concern would be centers/proceduralists who don't report this measure or participate in the appropriate registry. If this is a large number, or systematically biased (e.g. neuroradiologists) then public reporting of this measure might paint the wrong picture for the public.

- Threats to validity would be if/how risk adjustment is used. Recommend that no risk adjustment be performed for previously stated reasons and this is supported by the quality measure developers statements. Stroke/death are important quality indicators for carotid revascularization, especially for prophylactic operations. They will directly identify possible important meaningful differences that get at quality of care delivery for CAS.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- Data elements are generated and used by healthcare personnel during the provision of care and coded by someone other than the person obtaining the original information.
- All data elements are in defined fields with electronic clinical data
- The developer reports that in using VSGNE (which has been tracking stroke or death since 2005) and VQI, they have not had any difficulty obtaining these data.
- The developers note percent missing for this variable at less than 1%.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are there fees to belong to the registry?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- Requires individual assessment of medical records to ensure accurate diagnosis and capture of post-procedure stroke.
  How many CAS are performed by non-surgeons? Do they participate in the registries cited? If not, how would they report this measure? How would their results likely compare?

- Stroke/death are routinely reported and reliably abstracted from EHR.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure

- Centers for Medicare & Medicaid Services, Physician Quality Reporting System (PQRS) (measure #345)

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- PQRS is a reporting program that uses a combination of incentive payments and payment adjustments to promote reporting of quality information by eligible professionals (EPs). PQRS measures are used for public reporting on the Physician Compare website and for the quality component of the Value-Based Payment Modifier (VBPM).

Improvement results

- Although this measure is reported in PQRS, the developer did not provide these data.
- An analysis of 2010-2015 VQI self-reported data (175 centers, 544 providers, 3,342 procedures) found an outcome of 2.1% stroke or death within 30 days after a CAS procedure.
- The developer also notes a decrease in the percentage of cases with a reported death within 30 days of CAS has fallen from 2.1% to 1.6%, with the exception of in 2012 where the rate was 3.3%.

	Free from outcome	With Outcome
Overall, N=3342	3273 (97.9%)	69 (2.1%)
2010, n=100	98 (98%)	2 (2%)
2011, n=226	221 (97.8%)	5 (2.2%)
2012, n=549	531 (96.7%)	18 (3.3%)
2013, n=739	725 (98.1%)	14 (1.9%)
2014, n=749	735 (98.1%)	14 (1.9%)
2015, n=979	963 (98.4%)	19 (1.6%)

Unexpected findings (positive or negative) during implementation

- The developer notes that data definitions regarding asymptomatic status based on NASCET criteria have eliminated confusion about symptoms.
- The developer also notes they have not had any challenges with this measure.

Potential harms

- No potential harms reported.

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- Assuming that it can capture most providers of this procedure, it seems like a good measure, with high face validity for what the publis would like to know.

- The PQRS already provides a construct for public reporting and accountability for the proposed quality measure. The author of the measure did not provide any data from PQRS on these measures.

Criterion 5: Related and Competing Measures

Related/competing measures

- N/A

Harmonization

- N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1543 NQF Project: Surgery Endorsement Maintenance 2010

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

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

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process-health outcome; intermediate clinical outcome-health outcome)*:

discussed above

1c.2-3 Type of Evidence *(Check all that apply)*:

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

The combined endpoint of stroke/death is the accepted primary endpoint for both CAS and carotid endarterectomy. Variation in outcome has been established in randomized trials, cohort studies and meta analyses. This outcome measure has face validity among all providers of this service. Studies cited above have shown substantial variation in outcomes by provider when CEA is performed in asymptomatic patients. While such data does not yet exist for CAS, similar findings are expected due to the same patient population being treated.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*:

1c.6 Quality of Body of Evidence *(Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: Expert opinion.

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Stroke/death after CAS is the reporting standard recommended by the Society for Vascular Surgery.

1c.14 Summary of Controversy/Contradictory Evidence: The endpoint of stroke, death or myocardial infarction is a frequent endpoint in CAS studies. However, this is seldom used in CEA studies, and recent studies have shown that the impact of MI is much less than the impact of stroke after CAS. Thus, we favor stroke/death as the primary endpoint for this measure.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

1.) Carotid Artery Angioplasty and Stent Placement: Quality Improvement Guidelines to Ensure Stroke Risk Reduction, J Vasc Interv Radiol 2003;14;S317-9. 2.) Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis, JAMA 1995;273:1421-8. 3.) Management of Atherosclerotic Carotid Artery Disease: Clinical Practice Guidelines of the Society for Vascular Surgery, J Vasc Surg 2008;48:480-6. 4.) Clinical Competence Statement on Carotid Stenting: Training and Credentialing for Carotid Stenting-Multispecialty Consensus Recommendations, J Vasc Surg 2005;41:160-8. 5.) Percutaneous Transluminal Angioplasty and Stenting for Carotid Artery Stenosis, Cochrane Database Syst Rev 2007;(4):CD000515. 6.) Endarterectomy vs Stenting for Carotid Artery Stenosis: A Systematic Review and Meta-analysis, J Vasc Surg 2008;48:487-93. 7.) Carotid Stenting and Angioplasty, Circulation 1998;97:121-3. 8. Risk-adjusted 30-day outcomes of carotid stenting and endarterectomy: Results from the SVS Vascular Registry, J Vasc Surg 2008.

Added for 2016 Maintenance:

CAS Citations for evidence:

1. "Safety of stenting and endarterectomy by symptomatic status in the Carotid Revascularization Endarterectomy Versus Stenting Trial (CREST)." Silver FL(1), Mackey A, Clark WM, Brooks W, Timaran CH, Chiu D, Goldstein LB, Meschia JF, Ferguson RD, Moore WS, Howard G, Brott TG; CREST Investigators. Stroke. 2011 Mar;42(3):675-80. doi: 10.1161/STROKEAHA.110.610212.

2. "Randomized Trial of Stent versus Surgery for Asymptomatic Carotid Stenosis". Rosenfield K(1), Matsumura JS(1), Chaturvedi S(1), Riles T(1), Ansel GM(1), Metzger DC(1), Wechsler L(1), Jaff MR(1), Gray W(1); ACT I Investigators. N Engl J Med. 2016 Mar 17;374(11):1011-20. doi: 10.1056/NEJMoa1515706. Epub 2016 Feb 17.

3. "Experience matters more than specialty for carotid stenting outcomes" Sgroi, Michael D. et al. Journal of Vascular Surgery 2015 , Volume 61 , Issue 4 , 933 - 938.

4. Experience and outcomes with carotid artery stenting: an analysis of the CHOICE study (Carotid Stenting for High Surgical-Risk Patients; Evaluating Outcomes Through the Collection of Clinical Evidence).JACC Cardiovasc Interv. 2014 Nov;7(11):1307-17. doi: 10.1016/j.jcin.2014.05.027.

1c.16 Quote verbatim, <u>the specific guideline recommendation</u> *(Including guideline # and/or page #)*:

Presently there is no published guideline that places a threshold for acceptable stroke and death rates following CAS for the treatment of asymptomatic carotid stenosis. There is, however, an acceptable and published threshold of 3% for patients treated with the established surgical alternative, CEA. The AHA has determined that CEA in particular should only be

performed for asymptomatic carotid stenosis if the risk of the procedure was les than 3% stroke and/or death (2). It has been suggested that this is fairly generalizable to any form of intervention (1)

1c.17 Clinical Practice Guideline Citation: Risk-adjusted 30-day outcomes of carotid stenting and endarterectomy: Results from the SVS Vascular Registry, J Vasc Surg 2008.

1c.18 National Guideline Clearinghouse or other URL: NA

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: NA

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: NA

1c.24 Rationale for Using this Guideline Over Others:

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?

1c.25 Quantity: 1c.26 Quality: 1c.27 Consistency:

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1543_Evidence_MSF5.0_Data_-1-_2016.doc

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Better patient selection to avoid treating high risk patients who will likely experience stroke or death after CAS for asymptomatic patients which eliminates any benefit of the procedure.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Stroke or death following CAS has been the primary clinical endpoint for a number of clinical CAS trials. Stroke or death within 30 days following intervention is captured in the SVS Registry, Vascular Quality Initiative (VQI). This endpoint is easy to capture from claims data and registries. This outcome is particularly important for asymptomatic patients undergoing CAS, since this is a prophylactic procedure being proposed to prevent future stroke. Guidelines from the American Heart Association recommend CEA for such patients only if the risk of surgical death or stroke combined is less than 3%. While there is no similar level published as a guideline, the same clinical threshold of 3% can be used for asymptomatic patients undergoing CAS. Cochrane Database analysis of stroke or death within 30 days of CAS for asymptomatic carotid stenosis showed no difference between CEA and CAS in all patients as well for a subset of patients deemed "not suitable for surgery" (CEA). Similarly, two large industry-sponsored carotid stent trials, CAPTURE-2 and EXACT, both demonstrated outcomes for CAS in asymptomatic patients that were "comparable to those established

by the AHA for patients treated with CEA".

Stroke is defined as an acute neurological deficit due to an occlusive or hemorrhagic brain lesion that persists more than 24 hours. It can be substantiated by a new stroke seen on brain imaging, but this is not a requirement, i.e., clinical symptoms alone are sufficient. Both minor and major strokes will be counted, as long as the symptoms persist more than 24 hours. Stroke in either carotid distribution, or vertebrobasilar stroke is included, i.e., any postprocedural new neurologic deficit attributed to an occlusive or hemorrhagic brain lesion lasting more than 24 hours.

While stroke or death following CAS is an appropriate quality measure for either symptomatic or asymptomatic patients, we believe that the former group would require risk adjustment to allow fair comparisons, while we do not believe this is necessary for asymptomatic patients. For asymptomatic patients, it is incumbent upon the interventionalist to select only those patients of low periprocedural risk to benefit from CAS.

We propose that the denominator for this measure should be patients who have never been symptomatic in either the cerebral hemisphere ipsilateral to the carotid lesion, the contralateral hemisphere or the vertebrobasilar circulation(dizziness or lightheadedness alone are not considered symptoms). This group has the lowest risk of stroke with carotid intervention and also the lowest risk of stroke with medical therapy alone.

Adopting this outcome measure would likely have immediate impact on improving quality. Regional data have shown that feedback of the key outcome of stroke and death, in addition to some process measures after carotid endarterectomy reduced this outcome from 5.6% to 5.0% and in asymptomatic patients from 4.1% to 3.8%. The same is likely to hold true for CAS. Reporting time frame for hospitals should be on a yearly basis. The time frame for interventionalists should be cumulative over their career.

In an analysis of the VQI self-reported data for the time period of 2010 - 2015, across 175 centers with 544 providers reporting on 3,342 procedures, we found an outcome of 2.1% of a stroke or death within 30 days after a CAS procedure for all reported cases. And, with the exception of 2012 where it jumped to 3.3%, the percentage of cases with a reported death within 30 days of the CAS procedures has fallen,2% - 1.6%, even as the number of incidents has increased as the number of patients included in the denominator has increased.

In a review by center over the five year period, we found an interquartile range of 0% to 1.7% per center with the number of centers increasing each year.

While there has been some improvement over the last five years, there continues to be a performance gap regarding the number of deaths in this 2016 study.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
To date, there is no strong evidence that CAS for asymptomatic carotid stenosis provides a significant benefit to patients over best medical therapy. Nevertheless, CAS is being performed for the treatment of asymptomatic stenosis in multiple centers in the US. The results of controlled randomized trials are pending and should soon provide the Level 1 evidence required.

Although CAS is not approved for reimbursement by CMS for asymptomatic patients, this procedure is performed for asymptomatic patients in 65% of patients in VSGNE undergoing CAS. We suspect overuse in many of these patients.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Such data will become available if this measure is adopted for reporting and used by more centers with more varied population demographics than found in the New England region.

In our 2016 analysis of the 3,342 patients reported over the time frame of 2010 - 2015 to the SVS VQI, we found that the patients still experiencing a stroke or death within 30 days of a CAS procedure were older, had Medicare as their insurance, and were slightly more likely to be female. We did not see any specific differences related to race.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from**

the literature that addresses disparities in care on the specific focus of measurement. Include citations.
not available

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Percutaneous carotid intervention is a rapidly emerging field. Published trial results have established carotid stenting (CAS) in high risk surgical patients to be an effective alternative to carotid endarterectomy (CEA). It is well established that CEA benefits patients with asymptomatic >60% stenosis only if performed with a high degree of technical proficiency on appropriately selected patients. The same is proposed to hold true for CAS. This is particularly important when considering an asymptomatic population where the relative risk reduction with intervention is narrow when compared to medical management. Numerous publications have noted variation in the combined endpoint of stroke and death following carotid angioplasty and stent placement with embolic protection (5). Adoption of this outcome measure in the United States would likely disclose disperate results between hospitals and between providers, and lead to quality improvement when this information was provided to individual providers and participating centers. The SVS Vascular Registry has shown that outcome results are good for CAS, but variations exist between interventionalists and centers (8). Postoperative stroke or death is the accepted outcome parameter for this procedure, and its measurement and reporting would demonstrate variation and opportunity for improvemement. CAS is an elective procedure in nearly all cases. Patients can be referred or transferred to a center with the personnel and experience to perform this procedure with a high level of competence and any procedure that has "stroke" as a potential risk should be performed only by individuals with appropriate training and experience. (1)

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1.) Carotid Artery Angioplasty and Stent Placement: Quality Improvement Guidelines to Ensure Stroke Risk Reduction, J Vasc Interv Radiol 2003;14;S317-9. 2.) Executive Committee for the Asymptomatic Carotid Atherosclerosis Study. Endarterectomy for asymptomatic carotid artery stenosis, JAMA 1995;273:1421-8. 3.) Management of Atherosclerotic Carotid Artery Disease: Clinical Practice Guidelines of the Society for Vascular Surgery, J Vasc Surg 2008;48:480-6. 4.) Clinical Competence Statement on Carotid Stenting: Training and Credentialing for Carotid Stenting-Multispecialty Consensus Recommendations, J Vasc Surg 2005;41:160-8. 5.) Percutaneous Transluminal Angioplasty and Stenting for Carotid Artery Stenosis, Cochrane Database Syst Rev 2007;(4):CD000515. 6.) Endarterectomy vs Stenting for Carotid Artery Stenosis: A Systematic Review and Meta-analysis, J Vasc Surg 2008;48:487-93. 7.) Carotid Stenting and Angioplasty, Circulation 1998;97:121-3. 8. Risk-adjusted 30-day outcomes of carotid stenting and endarterectomy: Results from the SVS Vascular Registry, J Vasc Surg 2008.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

Quality Data Model (QDM).

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

De.6. Cross Cutting Areas *(check all the areas that apply):*
Safety : Complications

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.vascularqualityinitiative.org/wp-content/uploads/2016_PQRS_Information-v2-1.pdf

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

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** CAS defs v.01.09.doc

S.3. For endorsement maintenance, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
There are no changes since the last endorsement date.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year of their procedure who experience stroke or death during their hospitalization following elective carotid artery angioplasty and stent placement.

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Since hospitals have sufficient annual volume to generate accurate reporting levels, these are proposed for reporting every 12 months for hospital. Since surgeons have lower individual volume, we recommend annual reporting of the last 50 consecutive procedures, which may span more than one year, with suppression if < 10 procedures (ie, reported as too low volume to report).

S.6. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
ANY registry that includes hospitalization details and symptom status within 120 days is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Other registries that collect this same information could report these measures. Patients who were asymptomatic within one year of the CAS (CPT code 37215) who died or had a stroke recorded in the registry during that admission.ANY registry that includes hospitalization details and symptom status within 120 days is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Patients who were asymptomatic within one year of the CAS (CPT code 37215) who died or had a stroke recorded in the registry during that admission.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Patients over age 18 without preoperative carotid territory neurologic or retinal symptoms within one year immediately preceding carotid artery stenting.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
ANY registry that includes hospitalization details and symptom status within one year is required to identify patients for numerator inclusion. The Society for Vascular Surgery Vascular Quality Initiative (SVS VQI) and the Vascular Study Group of New England (VSGNE) are examples of registries that record such information, but the measure is not limited to these registries. Patients who were asymptomatic within one year of the CAS (CPT code 37215) are included.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Per PQRS Specifications for 2016:

DENOMINATOR EXCLUSIONS:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F
OR
Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Patients with NASCET criteria neurologic symptoms (transient ischemic attack, amaurosis, or stroke) within the one year immediately proceeding CAS.

DENOMINATOR EXCLUSIONS per PQRS 2016 specifications:
Symptomatic carotid stenosis: Ipsilateral carotid territory TIA or stroke less than 120 days prior to procedure: 9006F
OR
Other carotid stenosis: Ipsilateral TIA or stroke 120 days or greater prior to procedure or any prior contralateral carotid territory or vertebrobasilar TIA or stroke: 9007F

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Not required

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See "Scientific Acceptablility" section for rationale

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Number of asymptomatic patients undergoing CAS who have in hospital stroke or death / Number of asymptomatic patients undergoing CAS

INSTRUCTIONS:
This measure is to be reported each time a CAS is performed during the reporting period. It is anticipated that clinicians who provide services of CAS, as described in the measure, based on the services provided and the measure-specific denominator coding will report this measure. This measure may be reported by clinicians who perform the quality actions described in the measure based on the services provided and the measure-specific denominator coding.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Society for Vascular Surgery Vascular Quality Initiative Registry
Vascular Study Group of New England Registry

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting*

rules, or calculation of individual performance measures if not individually endorsed.)

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
1543_MeasureTesting_MSF5.0_Data_v1.doc

NATIONAL QUALITY FORUM

Measure missing data in MSF 6.5 from MSF 5.0

NQF #: 1543 NQF Project: Surgery Endorsement Maintenance 2010

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. (evaluation criteria)
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 field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

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)*:
A random sample of 100 patient records representing 5 procedures relevant to the measure from 5 different hospitals based on data collected during the past 2 years. In addition, in-hospital mortality was examined by claims based analysis of 7,205 patients discharged and recorded in the VSGNE registry between 2003 to 2007. These measures were originally tested in 2011 and this was the most recent data. All of the testing was approved by the NQF Steering Committee at the time that the measures were first approved in 2012. These measures are approved for PQRS reporting and working well. Regarding the sample and the data, this is an accepted testing practice to pull a sample for chart review to then compare to the data that was submitted to a registry.

2a2.2 Analytic Method *(Describe method of reliability testing & rationale)*:
A nurse abstractor completed a form based on medical record review for the variables relevant to this measure. The results of this chart review were then compared with the original registry data. The Kappa statistic was used to judge reliability of the data. For mortality validation, claims data from each of 12 hospitals were matched to patient identified data within the VSGNE registry to compare discharge status (alive vs. dead). Any discrepencies were then further evaluated based on a medical record audit.

2a2.3 Testing Results *(Reliability statistics, assessment of adequacy in the context of norms for the test conducted)*:
The key variables for this measure and testing results were:

1. Correct procedure (carotid artery stenting) performed. Kappa =1.0
2. Hospital mortality: Kappa = .91 (SE .01)
3. Hospital stroke: Kappa = 1.0
4. Asymptomatic 120 days pre-Rx: Kappa = .90 (SE .07)

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* are consistent with the evidence cited in support of the measure focus (*criterion 1c)* and identify any differences from the evidence:

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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 reliability

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)*:
Multiple sources from the medical record were used as the gold standard, and rates compared with literature. Please see the evidence listed in the NQF form under importance.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
The percentage of asymptomatic patients being treated in VSGNE of 60% corresponds to published data on this cohort. The postop stroke or death rate of 2.2% also correponds to published results for asymptomatic patients.

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 Data/Sample for analysis of exclusions *(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)*:
SVS Vascular Registry 805 asymptomatic patients undergoing elective CEA

2b3.2 Analytic Method (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)*:
measure calculation

2b3.3 Results *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
Death rate 2.0%, stroke rate 2.11% among 287 provider in 58 centers
Interquartile range was 0.3-8.6% for the combined endpoint

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 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 "Scientific Acceptablility" section for rationale. Risk adjustment is implicit within this quality measure as judged by the sponsor, the Society for Vascular Surgery, for the following reason. CAS in an asymptomatic patients is a prophylactic procedure designed to prevent future stroke. The decision to perform such a procedure requires the interventionist to calculate the patient´s risk-benefit ratio, in order to avoid post-CAS stroke or death that eliminate the benefit of the procedure. Risk adjustment based on patient factors should not be applied, since high risk patients should not undergo this prophylactic procedure, and using risk adjustment would reward interventionists who selected high risk patients for treatment.

2b4.2 Analytic Method (*Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)*:
N/A

2b4.3 Testing Results (*Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
N/A

2b4.4 If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:
N/A

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a*

sample, characteristics of the entities included):
see section 1.b.3 and above 2,d,5

2b5.2 Analytic Method *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)*:
Standard statistial analysis to determine 95% confidence interval for hospitals and providers to determine practical difference from mean

2b5.3 Results *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 Data/Sample *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)*:
no other data sources available

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: N/A

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
No disparities have been reported. Please see the new data under the importance sections of the NQF regular form that was required as part of the measure maintenance check list.

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

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

3a. Byproduct of Care Processes

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

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
generated by and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition,
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

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

In the VSGNE experience which has been tracking stroke or death as a major endpoint since 2005, we have not experienced any difficulty with obtaining data related to this endpoint. Our percent missing for this variable has been less than 1%. This has also been the case with the VQI.

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

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. **Current and Planned Use**

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

Planned	Current Use (for current use provide URL)
	Payment Program PQRS Approved Measure www.cms.hhs.gov

	Quality Improvement (Internal to the specific organization) Vascular Quality Initiative www.vascular.org

4a.1. **For each CURRENT use, checked above, provide:**

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

This measure is an approved measure for PQRS reporting. It is PQRS measure number 345. PQRS is the physician quality and payment program operated by the Centers for Medicare and Medicaid Services. It is a national program. We are not aware from CMS how may entities are reporting this measure nor how often it has been reported.

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

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Overall (N=3342) Free from outcome (N=3273, 97.9%) With outcome (N=69, 2.1%)

RATE BY YEAR

2010	100 (3%)	98 (98%)	2 (2%)
2011	226 (6.8%)	221 (97.8%)	5 (2.2%)
2012	549 (16.4%)	531 (96.7%)	18 (3.3%)
2013	739 (22.1%)	725 (98.1%)	14 (1.9%)
2014	749 (22.4%)	735 (98.1%)	14 (1.9%)
2015	979 (29.3%)	963 (98.4%)	16 (1.6%)

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the**

negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
Data definitions regarding asymptomatic status based on NASCET criteria have eliminated confusion about symtoms. Death is an accurate endpoint. Stroke has been accurately collected as judged by chart audits and comparison to claims data that has been done within VSGNE.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Society for Vascular Surgery
Co.2 **Point of Contact:** Sarah, Murphy, smurphy@vascularsociety.org, 312-334-2305-
Co.3 **Measure Developer if different from Measure Steward:** Society for Vascular Surgery
Co.4 **Point of Contact:** Jill, Rathbun, Jill_Rathbun@galileogrp.com, 703-217-7224-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
N/A

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2010
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 12, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

CAROTID ARTERY STENT DEFINITIONS (Include only carotid bifurcation or internal carotid artery stents) v.01.09

If more than one response applies, select the most severe (highest number) response for each data field.

Pre-op

Smoking: Prior = quit ≥ 1 year ago. Current = still smoking within last 12 months. Include cigarettes, pipe, or cigar.
HTN (Hypertension): Defined as ≥ 140/90, either systolic or diastolic, at admission or within last 6 months, or clearly documented in medical record.
Beta-blockers: Peri-operative = started w/in one month before surgery or during surgery. Chronic = >than one month before surgery.
Symptoms (Coronary artery disease**):** Stable angina = stable pattern or symptoms with or without antianginal medication. Unstable angina = new onset, increasing frequency, lasting > 20 min and/or rest angina.
CABG/PTCA: Coronary artery bypass, angioplasty, or stent.
CHF (Congestive Heart Failure): Documented CHF: Mild = SOB on exertion; Severe = SOB at rest, pulmonary edema, or pitting ankle edema. (Use 2 = mild if severity not documented.)
COPD: Not treated = COPD documented in record but not treated with medication. Meds include theophylline, aminophylline, inhalers or steroids
Dialysis: Transplant = patient has functioning kidney transplant; Dialysis = currently on hemo- or peritoneal dialysis.
Creatinine: Last available measurement taken before procedure. If multiple measurements, use highest within 30 days of surgery.
Stress Test: Includes stress EKG, stress echo, nuclear stress scans, within 2 years of surgery.
Pre-admin living: Use last living status before any current, acute hospitalization or rehab unit.
Previous Arterial:
Bypass - Any non-cardiac arterial bypass for occlusive disease
CEA - Carotid endarterectomy
Aneurysm Repair – Any known true arterial aneurysm repair (excluding cerebral or pseudo-aneurysm)
PTA/Stent – Of any non-cardiac artery
Major Amputation – Any amputation above the foot or hand
Pre-Op Medications: Taken within 36 hours of surgery. Statins include any HMG-CoA reductase inhibitor, such as Lipitor, Mevacor, Pravachol, Zocor, Lescol, etc. If Plavix is disontinued prior to surgery it should be coded = 0.
Pre-op Hemoglobin: Most recent pre-op hemoglobin within past 30 days.
Symptoms: Ocular: unilateral visual loss or major blurring, etc. Cortical: unilateral motor and/or memory loss, or dysphagia/aphasia, etc. Vertebrobasiliar: bilateral motor, sensory, or visual loss, diplopia, ataxaia, etc. Major cortical or vertebrobasilar stroke = disability causing non-independent living status. Minor stroke is non-disabling. Major ocular stroke = blindness, otherwise minor. Stroke<1 month means stroke within previous month before surgery, etc. TIA=transient ischemic attack completely resolved within 24 hours.
Non-specific: Not clearly a carotid or vertebrobasilar TIA, e.g., light-headedness, dizziness
Ipsilat stroke on CT/MRI: Carotid territory only.
Medical high risk: At least one factor required: > 80 years old, severe O2 dependent pulmonary disease, CHF w/in one month, or abnormal stress test.
Anatomic high risk: Previous endarterectomy, previous neck surgery or radiation, tracheal or pharyngeal stoma, lesion above C3, contralat laryngeal nerve palsy, or contralateral carotid occlusion.
Refused for surgery: Surgeon has evaluated patient and refuses to operate due to excessive risk.

ICA stenosis: Use most severe category by modality thought to be most accurate if multiple modalities used.

Procedure

Urgency: Urgent = surgery within 24 hrs of admit or patient can't be discharged; emergent = surgery within 6 hrs of admission.
Lesion length: Length of stenosis intended to be covered with stent.
Prophylactic Anti-bradyarrhythmic: Atropine or Glycopyrolate given prior to angioplasty
Pre-dilate before protection device: Angioplasty required in order to cross lesion with a protection device.
Proximal CCA stent: Stent placement in the origin of the CCA.
Bradyarrhythmia requiring tx: Any dose given post post-dilation.
Technical failure: Can't complete procedure – CAS procedure defined as starting with attempting to place long sheath into CCA.
Protection device failure: Can't cross lesion, filter clogged, difficulty removing filter, ICA spasm requiring treatment, neurological change during procedure.

Post-op

Cranial nerve injury: Any occurrence, transient or persisting: VII-facial droop or more severe; IX-swallowing difficulty unless other diagnosis confirmed; X- hoarseness unless laryngoscopy normal; XII-any tongue deviation or dis-coordination
Ipsilat/Contralat neurologic event: Cerebral or ocular. TIA = cortical or ocular symptoms <24hrs duration. Major cortical or vertebrobasilar stroke = disability causing non-independent living status. Otherwise, minor. Major ocular stroke = blindness, otherwise minor. Minor stroke is non-disabling.
Time of Onset Ipsila/Contralat: Time when first noticed, but if noted on awakening from anesthesia code as 1=intra-op. Use 2=≤6 hrs post-op if normal at completion of procedure, and then neurologic event developed.
2b3a Inhibitor: Integrilin, Aggrastat.
Reperfusion Symptoms: Seizures associated with headache, or hemorrhage on CT/MRI.
IV meds required: Indicates continuous infusion or more than one dose required more than one hour after surgery.
Myocardial Infarction: Troponin: by local standards for MI. EKG: new Q waves, new ST and T wave changes. Clinical: documentation of MI by clinical criteria or ECHO or other imaging modality.
Dysrhythmia: New rhythm disturbance requiring treatment with medications or cardio-version.
CHF: Pulmonary edema with requirement for monitoring or treatment in ICU.
Access site cx: Complications at puncture site. PA=pseudo-aneurysm.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1550
Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: The measure estimates a hospital-level risk-standardized complication rate (RSCR) associated with elective primary THA and TKA in Medicare Fee-For-Service beneficiaries who are 65 years and older. The outcome (complication) is defined as any one of the specified complications occurring from the date of index admission to 90 days post date of the index admission (the admission included in the measure cohort). The target population is patients 18 and over. CMS annually reports the measure for patients who are 65 years or older, are enrolled in fee-for-service (FFS) Medicare, and hospitalized in non-federal acute-care hospitals.
Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized complication rates (RSCRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA complications is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting complication rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. In addition, it has the potential to lower health care costs associated with complications.

Numerator Statement: The outcome for this measure is any complication occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. Complications are counted in the measure only if they occur during the index hospital admission or during a readmission. The complication outcome is a dichotomous (yes/no) outcome. If a patient experiences one or more of these complications in the applicable time period, the complication outcome for that patient is counted in the measure as a "yes".

Denominator Statement: The target population for the publically reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures. Additional details are provided in S.9 Denominator Details.
Denominator Exclusions: This measure excludes index admissions for patients:
1. Without at least 90 days post-discharge enrollment in FFS Medicare;
2. Who were discharged against medical advice (AMA); or,
3. Who had more than two THA/TKA procedure codes during the index hospitalization.

After applying these exclusion criteria, we randomly select one index admission for patients with multiple index admissions in a calendar year. We therefore exclude the other eligible index admissions in that year.

Measure Type: Outcome
Data Source: Administrative claims, Other, Paper Medical Records
Level of Analysis: Facility

Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

Maintenance of Endorsement -- Preliminary Analysis

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This outcome measure was initially endorsed in 2012 and calculates any complication occurring during the index admission to 90 days post-date of the index admission following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA).
- This diagram shows the potential actions that can affect the outcome.
- The developer submitted updated evidence for the measure noting the rates of complication and death following THA and TKA.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat vote on Evidence?*

Guidance from the Evidence Algorithm: Health outcome (Box 1) →relationship between outcome and at least one healthcare action identified/supported by stated rationale (Box 2) → Pass
Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

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

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

- Data submitted during the previous review show considerable variation , with an unadjusted mean of 4.98% , a range of 0 to 100% across 3, 311 hospitals in 2008. After adjustment for patient and clinical characteristics, the mean complication rate was 4.23%, ranging from 2.20 to 8.88%.
- Current performance data were analyzed from over 3,000 hospitals between 2011 and 2014. The median RSCR for the most recent reporting period was 3.1, the mean 3.2, and the range 1.4 to 6.9 demonstrating variation in complication rates.

	07/2011-06/2012	07/2012-06/2013	07/2013-06/2014	07-2011-06/2014
Number of hospitals	3,348	3,331	3,313	3,507
Number of admissions	291,250	295,222	305,983	892,455
Mean (SD)	3.4 (0.4)	3.1 (0.4)	3.0 (0.4)	3.2 (0.5)
Range	1.8 – 5.6	1.8 – 5.4	1.6 – 5.5	1.4 – 6.9
50th percentile	3.3	3.1	2.9	3.1

Disparities

- Disparities data show complication rates are similar among patients with social risk factors (dual eligible, African-Americans, and patients with AHRQ SES scores below 42.7) compared to patients without them. Rates tend to be higher among hospitals with higher proportions of patients with these risk factors.
- Note that data for dual eligible and African-American patients come from Medicare FFS claims; data for patients with an AHRQ SES index score below 42.7 come from Medicare FFS claims and the American Community Survey (2009-2013). AHRQ SES index scores describe the socioeconomic status of people living in defined geographic areas.

THA/TKA RSCRs by Characteristic July 2013 - June 2014

	Proportion of Dual Eligible		Proportion of African American Patients		Proportion of Patients with AHRQ SES Index Scores <42.7	
	Hospitals with Low Proportion	Hospitals with High Proportion	Hospitals with Low Proportion	Hospitals with High Proportion	Hospitals with Low Proportion	Hospitals with High Proportion
% of hospitals with characteristic	3.9%	11.7%	0.0%	6.3%	6.4%	24.0%
Number of Patients	319,189	105,920	94,924	220,523	262,967	135,873
Median	3.0	3.2	3.2	3.2	3.1	3.2

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*

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

Committee pre-evaluation comments

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

1a.

- Why just old folks? (because the light is good?) Why not include age in risk adjustment or stratify by age?
  Good list of outcomes, with about equal seriousness/risk
  Good data on reliability of ICD coding for these complications, especially with adjustment of infection definition
  "Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement." In measure notes from NQF. Is this fact well known?
  How many high and low outliers?
- This is a high volume and high cost procedure that is associated with relatively uncommon but serious complications. This submission looks to measure using a risk stratified methodology a composite of serious complications within 90 days of the index surgery. The relationship between the measured outcomes and service are clearly identified and support the rationale.
- Yes. The evidence shows small differences for events with a low incidence. Question: Is the contribution of statistical noise adequately separable when comparing individual hospitals as opposed to classes of hospitals? The question is made in the context of the measure now being collected for application to the Value Based Purchasing program in 2019 and currently is being used to distinguish by percentile differences quality in the CJR (THA/TKA bundle required of 23% of US hospitals).

1b. Performance Gap

- National performance data was presented that shows variation in outcomes that warrant investigation of process improvement. Disparities were evaluated and there is some evidence that SES is a contributor to worse outcomes. How exactly it influences outcomes is unclear.
- The performance gap is the variability in rates of complications. The variability is adequate.
  In terms of potential disparities, there were more complications in the hospitals with higher proportions of African-American patients, those with dual eligibility, and those patients with AHRQ SES scores less than 42.7. These results were discounted as being more affected by the treating hospital than the patients.
  My question for the stewards is based on the following. The median performance hospitals showed that race alone does not drive the risk.If one accepts that the race in and of itself does not add risk, the race distinction can be, unfortunately in our society, a surrogate for poverty, in particular in urban areas. Just because the other patients at that hospital are not African Americans, might not have dual eligibility, and might be just above the cut-off for the SES designation, it does not mean that they are “rich”. In fact, they are more likely to be living in varying degrees of poverty. The neighborhoods that such a hospital cares for is overall more likely to be caring for a spectrum of poverty, that if averaged across all patients, would score a lower SES than hospitals with lower proportions of such patient classes.

  Social and economic dysfunction of the poorest communities would ordinarily correlate with less health care sophistication, poorer life choices, higher degrees of alcohol and drug use, less access to care and transportation, and living environments that are not ideal. This would be the common denominator across the community at large from in which the population subsets of minorities, dual eligibility, and those patients with SES scores less than 42.7 live.

  If the overall composition of the hospital’s population is not taken into consideration as a risk factor for performance, is the hospital effect being “double counted” in a negative way? This is a testable question; if the above argument is true, the distribution of SES scores across the entirety of the hospital population should be lower in hospitals with higher proportion of AA, DE, and the SES cut-off score. It is the relative poverty of the entire patient population, not just those that can be defined by dichotomous factors including those patients with an SES score below 42.7. In effect, the very best hospital in an impoverished

urban environment might be encumbered with social obstacles affecting all patients in ways that a suburban hospital does not encounters; it is the community served as a whole, not just the selected risk factors, that could need risk adjustment and provide disparities.

- The measure performance is consistent over the time periods evaluated in the submission. The model design is similar to other measurement models currently in use.
- Multiple complications are captured. Weighting of the impact on the patient for the individual complications is not provided.

  It should be pointed out that vascular complications/amputations are not captured as codes. Peripheral neurologic injury is also not captured. Fortunately both are rare events, but would more important if the complications were weighted in terms of impact

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications **Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. **Data source(s):** The developer lists Administrative Claims Paper Medical Records, and Census Data/American Community survey as data sources. **Specifications:** • This measure is specified as a facility level measure for the hospital/acute care setting. • The numerator includes patients with any complication from THA/TKA procedure occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. The outcome is dichotomous (yes or no). Patients experiencing one or more of the specified complications in the applicable time period is counted as yes for this measure. o The following complications are counted as during the index admission or within a specified time frame from the date of the index admission : AMI, pneumonia, sepsis/septicemia/shock (within 7 days); surgical site bleeding, pulmonary embolism, or death (within 30 days); mechanical complication or periprosthetic joint infection/wound infection (within 90 days). • The developer notes that the outcome is any of the specified complications occurring during the index admission or during a readmission. The specific time frame varies by complication. • The denominator includes Medicare FFS beneficiaries at least 65 years of age undergoing THA/TKA. • ICD-9 and 10 codes and a crosswalk are included in the data dictionary. • Exclusions listed are index admissions for patients without at least 90 days post-discharge enrollment in FFS Medicare; patients discharged against medical advice; and patients who had more than two THA/TKA procedure codes during the index hospitalization. The developer notes that patients with more than two THA/TKA are rare and could be due to a coding error, hence the exclusion. • This outcome measure is risk-adjusted using a statistical risk model with 33 risk factors. • This measure is calculated as the ratio of the number of predicted to the number of expected admissions with a complication, multiplied by the national observed complication rate. • No updates were made to specifications of this measure for 2016 public reporting but a summary of all

measure specification updates from 2013 to 2015 are found here.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Summary of the reliability testing from the prior review:

- The analysis submitted for reliability testing for the previous evaluation demonstrated similar risk model performance using development and validation samples. This testing does not meet current NQF requirements for reliability testing.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The dataset used for testing included Medicare Parts A and B claims and Medicare Enrollment Database from April 1, 2011 – March 30, 2014. Within the dataset were 892,455 index admissions and 3,507 hospitals.
- Developers used a split-sample methodology to test the measure score reliability. The dataset was split into two samples and tested by randomly selecting 50% Medicare patients aged 65 years and over in the most recent three year cohort. The level of agreement between scores was compared using the intra-class-correlation coefficient (ICC). The ICC demonstrates the percentage of variance in score results that is due to true or real variance between hospitals.
- Although the developer reported assessing data element reliability by comparing model variable frequencies and odds ratios from logistic regression models across the most recent three years of data, NQF does not consider temporal consistency to be a valid method of demonstrating reliability of data elements.

Results of reliability testing

- Measure score reliability results
  - Of 892,455 admissions, 445,352 were index admissions from 2,826 hospitals in one sample. The other sample included 447,103 admissions from 2,851 hospitals. The ICC was 0.45, indicating that 45% of the variance in scores are due to differences between hospitals. According to the Landis and Koch classification, an ICC value of 45% can be interpreted as moderate agreement. However, a value of 0.7 is often regarded as a minimum acceptable reliability value.
  - Although the developer notes that the analysis is limited to hospitals with 12 or more cases in each split sample, the measure is not specified to include a minimum data sample of 12 cases.
  - The ICC is based on a split sample of three years data, resulting in a volume of patients in each sample to 1.5 years of data whereas the measure is reported with the full three years of data.

Questions for the Committee:

- o *Is the test sample adequate to generalize for widespread implementation?*
- o *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*
- o *Do the results demonstrate meaningful differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empiric reliability testing (Box 2) →performance score testing (Box 4) →appropriate method of testing (Box 5) → moderate certainty of reliability (box 6b)

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- o *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Summary of the validity testing from the prior review:
The developer did not include updates to validity testing for this measure. A summary of testing from the previous evaluation is provided below.

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method

- Validation of claims-based definition of complications consisted of comparing complications (or no complication) coded in the claims to what was documented in the medical record. Eight hospitals, representing 644 patients were included .Developers calculated overall measure agreement by comparing exact codes.
- The developer states that measure validity is also demonstrated through use of established measure development guidelines, and by systematic assessment of measure face validity by a technical advisory panel. NQF does not consider technical panel review an appropriate method of demonstrating face validity.

Validity testing results

- Results of validation of claims-based definition of complications showed overall agreement of 93% (598/644 patients) between complications in claims and abstracted medical record results. Developers

then examined overall agreement in patients with and without complications, finding initial agreement at 86% for patients with a complication and 99% for those without a complication.

- Developers proposed removal of ICD 9 code for 'other postoperative infection' since it was not specific enough to sepsis. They also proposed combining 'wound infection' and 'periprosthetic joint infection' as a single complication. Measure agreement then improved to 99% (635/644).

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- To determine the impact of exclusions on the cohort, overall frequencies and proportions were calculated for the total cohort excluded for each exclusion criterion.
  - patients discharged against medical advice – 0.01%
  - patients without at least 90 days post discharge enrollment in FFS Medicare – 0.15%
  - patients admitted for the index procedure and transferred to another facility – 0.20%
  - patients with more than two THA/TKA procedure codes during the index hospitalization - <0.01%
- Results showed that none of these exclusions were likely to affect the measure score.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment model: ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**
SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

Description of the model

- This measure is risk-adjusted using the hierarchical logistic regression model with 33 factors to create a hospital-level 90-day RSCR. This approach accounts for variance in outcomes within and between hospitals.
  - The model adjusts for age (65 and older), male gender, index with admissions with an elective THA procedure, number of procedures performed, and several clinical risk factors.
  - Only comorbidities that conveyed information about the patient at that time or in the 12 months prior, and not complications that arose during the course of the admission, were included in the risk adjustment.
- Candidate variables were patient level risk adjustors predictive of complications, based on empirical analysis, prior literature, and clinical judgement. For each patient, covariates were obtained from Medicare claims extending 12 months prior to and including the index admission. The model adjusted for case differences based on the clinical status of the patient at the time of admission using condition categories. The final model includes 33 factors showing strong association with complications.

Performance of the model

Discrimination statistics:

- The c-statistic reflects how accurately a statistical model distinguishes between a patient with and without an outcome. C-statistic values range from 0.5 to 1.0, where a value of 0.5 indicates the model is no better than chance at making a prediction of patients with and without the outcome of interest.
- **Dataset 1** includes the 2015 public reporting cohort: Medicare Part A inpatient and outpatient claims, Part B claims, and Medicare Enrollment Database and **Dataset 3** includes 2008 Medicare Part A inpatient and outpatient claims, and Part B outpatient claims.
    - The c-statistic for the first half of Dataset 3 sample (development sample) was 0.69 . (lowest decile 2%, highest decile 15%).
    - The c-statistic for the second half of Dataset 3 sample (validation sample) was 0.70 (lowest decile 2%, highest decile 15%).
    - The c-statistic for Dataset 1 (current public reporting data) was 1 (lowest decile 1.5%, highest decile 7.4%).
- C-statistics for Datasets 1 and 3 indicate consistent and good model discrimination. The wide range between the lowest and highest deciles indicate the ability of the model to distinguish high-risk subjects from low-risk patients

Calibration statistics:

- Calibration statistics for Dataset 3 were close to 0 and close to 1 at either end indicating good calibration of the model.

Risk Decile Plot

- The developer notes that the risk decile plot for Dataset 1demonstrated excellent discrimination of the model and good predictive ability since higher deciles of the predicted outcomes are associated with higher observed outcomes.

Overall

- Interpreting the three diagnostic results together, the developer states that the risk adjustment model adequately controls for differences in patient characteristics.

Conceptual basis and empirical support for potential inclusion of SDS factors in risk-adjustment approach

- The developer noted that although recent literature evaluates the relationship between SES or race and postoperative complications our outcomes, few studies directly address causal pathways or examine the role of the hospital in these pathways. Additionally, there is no clear consensus on which risk factors demonstrate the strongest relationships with complications. Three domains of SES factors that have been examined in the literature include patient level variables, neighborhood/community level variables, and hospital-level variables.
- The developer identified 4 conceptual pathways to consider:
    - Relationship of SES factors or race to health at admission
    - Use of low quality hospitals – The developer states that patients of lower income, lower education or unstable housing have been shown to hot have access to high quality facilities since these facilities are less likely to located in areas with large populations of poor patients
    - Differential care within a hospital – The developer states that African-American patients may experience differential, lower quality or discriminatory care; patients of lower education may also require differentiated care that clients may not necessarily receive
    - Influence of SES on complication risk outside of hospital quality and health status
- Based on the interpretation of the literature and analysis of the conceptual pathways, three SES and race variables were considered. Analyses of the strength and significance of the variables in a multivariable model found the effect size of each of the variables to be moderate, with the c-statistics nearly unchanged. Additionally, the inclusion of these variables in the model had little to no effect on hospital performance.
    - Factor – Effect size, median absolute change in hospital RSCR
        - Dual-eligible status – OR 1.21, 0.052%
        - African-American race – OR 1.07, 0.0253%

- AHRQ SES Index – OR 1.07, 0.0237%
- Decomposition analysis found that both dual eligible and low AHRQ SES Index (equal to or below 42.7) effects were significantly associated with hip/knee complications. The developer states that the inclusion of dual eligible or low AHRQ SES index variables in the model would adjust for both patient level differences and differences between hospitals, which could hide signs of hospital quality.
- Based on these results the developer decided not to include any of the SDS factors analyzed in the final-risk adjustment model.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Analyses of Medicare FFS data show variation in RSCR among hospitals using data from Dataset 1 (April 2011 – March 2014) with a median RSCR of 3.1% and ranging from 1.4% to 6.9%. The interquartile range was 2.9% to 3.4%. The developer notes that although there has been a downward trend in complication rates between 2011 and 2014, the complication rate following THA/TKA remains higher than expected for an elective procedure at 2.95% in 2013-2014, suggesting meaning differences across hospitals.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable.

2b7. Missing Data

- The developers did not assess the frequency of missing data, stating that missing values are rare since the measure uses claims data.

Guidance from Validity Algorithm: Precise specifications (Box 1) → potential threats to validity assessed (Box 2) → empirical validity testing (Box 3) →measure score testing (Box 6) → method conceptually sound conceptualization (Box 7) → moderate confidence that scores are a valid indicator of quality (Box 8b)

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications: Reliability-Specifications

- The developers really need to address in the meeting the reliability testing. They are proposing a facility based score yet the interclass-correlation coefficient is 0.45. This suggests that only 45% of the variation should be attributed to the facilities. Previously the committee has recommended such coefficients to be in the 0.6 or above range. This is my BIGGEST concern about the utility of this measure.

- The administrative data set source of data precludes certain orthopedic specific risk factors. There is

preliminary data that 7-10 such factors could raise the c-stastic of 0.65 to above 0.7.

2a2. Reliability – Testing

- see above comments.
- The measure has a moderate ICC score of 45%. The C-statisitc is only 0.65.

2b.1 Validity – Specifications:

- It has face validity but may not be able to appropriately attribute accountability which is a necessary requirement of meaningful quality improvement and public reporting.
- This raises the question of tiered validity depending on the end-use. The original endorsement was found to be valid by it's TEP and the NQF for public reporting using three categories of below expected, expected, and above expected levels of performance. It is now being used for VBP penalties starting in 2019 and for eligibility for reward and price point setting within the CJR. These report the results as percentile performance intervals of small differences across a small range. Does the face validity ascribed to the measure from the TEP apply to these new uses?

2b2. Validity – Testing

- There is very good agreement between the claims based complications and abstracted clinical data in a test set. The model is based upon ICD-9 coding. The developers have submitted a ICD-9 to ICD-10 crosswalk. They report that they continue to refine this.
- The validity of the data source needs discussion given the new end-uses that are trying to making much finer discernments.

  The reference for their validation study is not given in this application. It has been referenced in the past, and can be found at the end of one of the versions of version 2 of the measure (starting on page 72). The measure intends to capture the number of complications. The study to assess the validity of the administrative data compared to chart review reported overall accuracy rates that were true complications + true no complications/total. Numerically, 319 complications were captured through the administrative data. Those patients were added to 325 patients without captured complications. None of the 325 without a complication recorded were in discordance with the charts. Of the 319 patients with a complication, however, there were 86 patients with 97 discrepancies. By changing some of the complication definitions and codes, the discrepancy rate was lowered to what was considered a “true” number of discrepancies, which was 30 out of 319. Combining those numbers with the 325 patients without captured complications caused the authors to then say that the administrative data was valid with an agreement of 99%.

  The complication rates are low with a median around 4.2%. Patients will not have a complication roughly 96% of the time. The concern is that the numerator for the measure is the number of complications captured, regardless of how many were “missed” in the total. The question is: after the coding corrections, isn’t it the 30/319 (10%) non-agreement of complications that is the test of the validity? Given that the rare complication diagnoses that were removed from the complete set of rare diagnoses, was a second validity study in order with the new model? Finally, given that the measure is assessing the differences between hospitals in terms of percentile performances that is measured in tenths of a percentage point, is the data base accurate enough when there is, at best, a 10% discrepancy between chart and administrative data?

2b3-7. Threats to Validity

- NO major issues identified.
- There are orthopedic specific risk factors that have been reported as raising the c-statistic. The capture/delivery of that data is an issue.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

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

- All data elements are in defined fields in electronic claims and generated or collected by and used by health care personnel during the provision of care. The data are coded by someone other than the person obtaining original information.
- Administrative data are routinely collected as part of the billing process.
- There are no fees associated with the use of the measure.

Questions for the Committee:

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

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- All the data should be readily available
- The addition of orthopedic specific risk factors, if pursued, would require coding options.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

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

Accountability program details

- This measure is currently used in the CMS Hospital Inpatient Quality Reporting Program.
- The developer notes that for the 2015 public reporting period, the RSCR was reported for 3,507 hospitals and included 892,455 admissions.

Improvement results

- The developer reports progress in reducing the RSCR following THA/TKA. The median RCRR decreased by 0.4 absolute percentage points from April 2011-March 2012 (median RSCR: 3.3%) to April 2013-March 2014 (median RSCR: 2.9%). The median hospital RSCR from April 2011-March 2014 was 3.1% (IQR 2.9% - 3.4%).

Unexpected findings (positive or negative) during implementation

- The developer notes there are no unexpected findings to report.

Potential harms

- The developer did not identify any unintended consequences during measure development, model testing, or re-specification.

***Questions for the Committee**:*

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

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- I have concerns about the low ICC which translates into a major usability issue.
- The original end-use was for public reporting. That has been extended to potential penalties under VBP in 2019 (being collected currently). It is also now being used as one of the quality metrics being used to score performance in the CJR bundle that affects the discount price setting as well as eligibility for reward. It has also been accepted through a consortium decision as one of the quality measures that would be used in extending bundled plans to beyond FFS Medicare, as reported in the HCPLAN white paper.

  Unintended consequences: The measure is being used in a way that causes hospitals to compete in a zero sum environment of winners an losers. The hospitals need to expect that the results will improve due to that competition. With the relatively weak risk adjustment provided, part of that competition will potentially be pursued through risk shedding of patient condition classes perceived as having higher risk. Recent professional society polling shows that thresholds for reversible risk are being widely made using varied parameters and many are doing so because of the concerns of reporting and payment penalties. The act of active selection of only low risk patients ('cherry picking') is a risk in terms of patient access to care, even for those patients with irreversible risk factors. Although anecdotal, all surgeons know

providers/hospitals that are disproportionally more selective than average. Such perceptions of risk can be extended to social risk factors

Those hospitals who have higher percentages of patients from more challenged economic classes face potential penalties disproportional to what might be equal quality of care compared to those serving more wealthy communities.

Ideally, hospitals would be monitored for changes in case complexity and/or patient demographics.

The last potential unintended consequence would be that the zero sum competition might induce a lowering of diagnostic intensity for discovery of complications (possibly fewer CTA's, less cultures, less aggressive treatment of impending infections, etc.).

Criterion 5: Related and Competing Measures

Related or competing measures

- 0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB).
- 0564 : Cataracts: Complications within 30 Days Following Cataract Surgery Requiring Additional Surgical Procedures
- 1551 : Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
- 2052 : Reduction of Complications through the use of Cystoscopy during Surgery for Stress Urinary Incontinence

Harmonization

The developer reports this measure is harmonized.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1550
Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 5/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

☒ Health outcome: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process: Click here to name the process

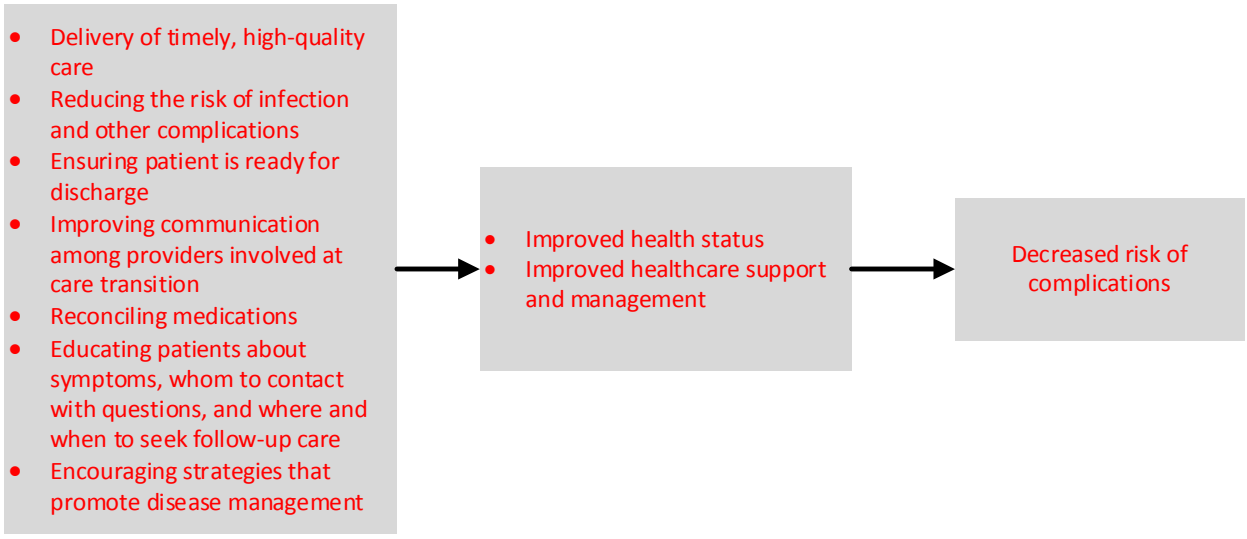
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.



The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized complication rates following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). Measurement of patient outcomes allows for a broader view of a hospital's quality of care that encompasses more than what can be captured by individual process of care measures. More specifically, complex and critical aspects of care, such as communication between providers, prevention of, and response to complications, patient safety and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This complication measure was developed to identify institutions, whose performance is better or worse than expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about the quality of care.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

In 2010 there were 168,000 THAs and 385,000 TKAs performed on Medicare beneficiaries 65 years and older (National Center for Health Statistics, 2010). Although these procedures dramatically improve quality of life, they are costly. In 2005, annual hospital charges totaled $3.95 billion and $7.42 billion for primary THA and TKA, respectively (Kurtz et al., 2007). These costs are projected to increase by 340% to 17.4 billion for THA and by 450% to 40.8 billion for TKA by 2015 (Kurtz et al., 2007). Medicare is the single largest payer for these procedures, covering approximately two-thirds of all THAs and TKAs performed in the US (Ong et al., 2006). Combined, THA and TKA procedures account for the largest procedural cost in the Medicare budget (Bozic et al., 2008).

Since THAs and TKAs are commonly performed and costly procedures, it is imperative to address quality of care. Complications increase costs associated with THA and TKA and affect the quality, and potentially quantity, of life for patients. Although complications following elective THA and TKA are rare, the results can be devastating. Rates for periprosthetic joint infection following THA and TKA range from 1.6% to 2.3%, depending upon the population (Bongartz et al., 2008; Kurtz et al., 2010). Reported 90-day death rates following THA range from 0.7% (Soohoo et al., 2010) to 2.7% (Cram et al., 2007). Rates for pulmonary embolism following TKA range from 0.5% to 0.9% (Cram et al., 2007; Mahomed et al., 2003; Khatod et al., 2008; Solomon et al., 2006). Rates for wound infection in Medicare population-based studies vary between 0.3% and 1.0% (Cram et al., 2007; Mahomed et al., 2003; Solomon et al., 2006). Rates for septicemia range from 0.1%, during the index admission (Browne et al., 2010) to 0.3%, 90 days following discharge for primary TKA (Cram et al., 2007). Rates for bleeding and hematoma following TKA range from 0.94% (Browne et al., 2010) to 1.7% (Huddleston et al., 2009).

The variation in complication rates across hospitals indicates there is room for quality improvement and targeted efforts to reduce these complications could result in better patient care and potential cost savings.

Measurement of patient outcomes allows for a comprehensive view of quality of care that reflects complex aspects of care such as communication between providers and coordinated transitions to the outpatient environment. These aspects are critical to patient outcomes, and are broader than what can be captured by individual process of care measures.

The THA/TKA hospital-specific risk-standardized complication rate (RSCR) measure is thus intended to inform quality-of-care improvement efforts, as individual process-based performance measures cannot encompass all the complex and critical aspects of care within a hospital that contribute to patient outcomes.

References:

Bongartz T, Halligan CS, Osmon DR, et al. Incidence and risk factors of prosthetic joint infection after total hip or knee replacement in patients with rheumatoid arthritis. Arthritis Rheum. Dec 15 2008;59(12):1713-1720.

Bozic KJ, Rubash HE, Sculco TP, Berry DJ. An analysis of medicare payment policy for total joint arthroplasty. *J Arthroplasty.* Sep 2008;23(6 Suppl 1):133-138.

Browne J, Cook C, Hofmann A, Bolognesi M. Postoperative morbidity and mortality following total knee arthroplasty with computer navigation. Knee. Mar 2010;17(2):152-156.

Cram P, Vaughan-Sarrazin MS, Wolf B, Katz JN, Rosenthal GE. A comparison of total hip and knee replacement in specialty and general hospitals. J Bone Joint Surg Am. Aug 2007;89(8):1675-1684.

Huddleston JI, Maloney WJ, Wang Y, Verzier N, Hunt DR, Herndon JH. Adverse Events After Total Knee Arthroplasty: A National Medicare Study. The Journal of Arthroplasty. 2009;24(6, Supplement 1):95-100.

Khatod M, Inacio M, Paxton EW, et al. Knee replacement: epidemiology, outcomes, and trends in Southern California: 17,080 replacements from 1995 through 2004. Acta Orthop. Dec 2008;79(6):812-819.

Kurtz S, Ong K, Lau E, Bozic K, Berry D, Parvizi J. Prosthetic joint infection risk after TKA in the Medicare population. Clin Orthop Relat Res. 2010;468:5.

Kurtz SM, Ong KL, Schmier J, et al. Future clinical and economic impact of revision total hip and knee arthroplasty. J Bone Joint Surg Am. Oct 2007;89 Suppl 3:144-151.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. J Bone Joint Surg Am. Jan 2003;85-A(1):27-32.

National Center for Health Statistics. National Hospital Discharge Survey: 2010 table, Procedures by selected patient characteristics - Number by procedure category and age. Available at http://www.cdc.gov/nchs/data/nhds/4procedures/2010pro4_numberprocedureage.pdf.

Ong KL, Mowat FS, Chan N, Lau E, Halpern MT, Kurtz SM. Economic burden of revision hip and knee arthroplasty in Medicare enrollees. Clin Orthop Relat Res. May 2006;446:22-28.

Solomon DH, Chibnik LB, Losina E, et al. Development of a preliminary index that predicts adverse events after total knee replacement. Arthritis & Rheumatism. 2006;54(5):1536-1542.

Soohoo NF, Farng E, Lieberman JR, Chambers L, Zingmond DS. Factors That Predict Short-term Complication Rates After Total Hip Arthroplasty. Clin Orthop Relat Res. Sep 2010;468(9):2363-2371.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***
☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***
☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***
☐ Other – ***complete section 1a.8***

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

N/A

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

N/A

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

N/A

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

N/A

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

N/A

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

N/A

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

N/A

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

N/A

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

N/A

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

N/A

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

N/A

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

N/A

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

N/A

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

N/A

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

N/A

_____

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
NQF_1550_HipKnee_Complication_NQF_Evidence_Attachment_v1.0.docx

1b. Performance Gap

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

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

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

The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized complication rates (RSCRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA complications is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting complication rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. In addition, it has the potential to lower health care costs associated with complications.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We examine the distribution of hospital performance scores to demonstrate the current gap in quality among measured hospitals. The results below indicate that the median RSCR for all measured hospitals in the most recent 3-year reporting period was 3.1. The mean was 3.2 and the range was 1.4 to 6.9 showing persistent variation in complication rates across hospitals.

Distribution of Hospital THA/TKA RSCRs over Different Time Periods
Results for each data year
Characteristic//07/2011-06/2012//07/2012-06/2013//07/2013-06/2014//07-2011-06/2014
Number of Hospitals// 3,348 // 3,331 // 3,313 // 3,507
Number of Admissions// 291,250 // 295,222 // 305,983 // 892,455
Mean (SD)// 3.4 (0.4) // 3.1 (0.4) // 3.0 (0.4) // 3.2 (0.5)
Range (min. – max.)// 1.8 – 5.6 // 1.8 – 5.4 // 1.6 – 5.5 // 1.4 – 6.9
Minimum// 1.8 // 1.8 // 1.6 // 1.4
10th percentile// 2.9 // 2.7 // 2.5 // 2.5
20th percentile// 3.1 // 2.9 // 2.7 // 2.8
30th percentile// 3.2 // 3.0 // 2.8 // 2.9
40th percentile// 3.3 // 3.1 // 2.9 // 3.0
50th percentile// 3.3 // 3.1 // 2.9 // 3.1
60th percentile// 3.4 // 3.1 // 3.0 // 3.2
70th percentile// 3.5 // 3.3 // 3.1 // 3.4
80th percentile// 3.6 // 3.4 // 3.2 // 3.5
90th percentile// 3.9 // 3.6 // 3.5 // 3.9
Maximum// 5.6 // 5.4 // 5.5 // 6.9

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

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

The results of this analysis show that complication rates are similar among patients with social risk factors compared to patients without these risk factors, although they tend to be slightly higher among hospitals with higher proportions of patients with these risk factors. However, the difference in median complication rates ranges from 0.0 to 0.2 absolute percentage points depending upon the social risk factor measure, indicating relatively small differences across groups.

Distribution of THA/TKA RSCRs by Proportion of Dual Eligible Patients:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims
Characteristic//Hospitals with a low proportion (=3.9%) Dual Eligible patients // Hospitals with a high proportion (=11.7%)
Dual Eligible patients Number of Measured Hospitals// 703 // 702
Number of Patients// 319,189 patients in low-proportion hospitals // 105,920 in high-proportion hospitals
Maximum// 5.1 // 5.5
90th percentile// 3.8 // 4.1
75th percentile// 3.4 // 3.6
Median (50th percentile)// 3.0 // 3.2
25th percentile// 2.7 // 2.9
10th percentile// 2.4 // 2.7
Minimum // 1.4 // 2.0

Distribution of THA/TKA RSCRs by Proportion of African-American Patients:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims
Characteristic// Hospitals with a low proportion (=0.0%) African-American patients // Hospitals with a high proportion (=6.3%) African-American patients
Number of Measured Hospitals// 712 // 703
Number of Patients// 94,924 patients in low-proportion hospitals // 220,523 patients in high-proportion hospitals
Maximum// 4.9 // 5.5
90th percentile// 3.9 // 4.1
75th percentile// 3.5 // 3.6
Median (50th percentile)// 3.2 // 3.2
25th percentile// 2.9 // 2.9
10th percentile// 2.7 // 2.6
Minimum // 2.0 // 1.8

Distribution of THA/TKA RSCRs by Proportion of Patients with AHRQ SES Index Scores Below 42.7:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims and the American Community Survey (2009-2013) data
Characteristic//Hospitals with a low proportion of patients below AHRQ SES index score of 42.7 (=6.4%) // Hospitals with a high proportion of patients below AHRQ SES index score of 42.7 (=24.0%)
Number of Measures Hospitals// 702 // 703
Number of Patients// 262,967 patients in hospitals with low proportion of patients below AHRQ SES index score of 42.7 // 135,873 patients in hospitals with high proportion of patients below AHRQ SES index score of 42.7
Maximum// 6.9 // 5.5
90th percentile// 3.9 // 4.0
75th percentile// 3.4 // 3.5
Median (50th percentile)// 3.1 // 3.2

25th percentile// 2.8 // 2.9
10th percentile// 2.5 // 2.6
Minimum // 1.4 // 2.0

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
THA and TKA are commonly performed procedures that improve quality of life. In 2003, there were 202,500 THAs and 402,100 TKAs performed (Kurtz et al., 2007a) and the number of procedures performed has increased steadily over the past decade (Kurtz et al., 2007b, Ong et al., 2006).

Although these procedures dramatically improve quality of life, they are costly. In 2005, annual hospital charges totaled $3.95 billion and $7.42 billion for primary THA and TKA, respectively (Kurtz et al., 2007b). These costs are projected to increase by 340% to $17.4 billion for THA and by 450% to $40.8 billion for TKA by 2015 (Kurtz et al., 2007b). Medicare is the single largest payer for these procedures, covering approximately two-thirds of all THAs and TKAs performed in the US (Ong et al., 2006). Combined, THA and TKA procedures account for the largest procedural cost in the Medicare budget (Bozic et al., 2008).

Because these are commonly performed and costly procedures, it is imperative to address quality of care. Complications increase costs associated with THA and TKA and affect the quality, and potentially quantity, of life for patients. Although complications following elective THA and TKA are rare, the results can be devastating. Rates for periprosthetic joint infection following THA and TKA range from 1.6% to 2.3%, depending upon the population (Bongartz et al., 2008; Kurtz et al., 2010). Reported 90-day death rates following THA range from 0.7% to 2.7% (Soohoo et al., 2010; Cram et al., 2007). Rates for pulmonary embolism following TKA range from 0.5% to 0.9% (Cram et al., 2007; Mahomed et al., 2003; Khatod et al., 2008; Solomon et al., 2006). Rates for wound infection in Medicare population-based studies vary between 0.3% and 1.0% (Cram et al., 2007; Mahomed et al., 2003; Solomon et al., 2006). Rates for septicemia range from 0.1%, during the index admission to 0.3%, 90 days following discharge for primary TKA (Browne et al., 2010; Cram et al., 2007). Rates for bleeding and hematoma following TKA range from 0.94% to 1.7% (Browne et al., 2010; Huddleston et al, 2009).

Furthermore, hospitals vary in their rate of complications. Analyses in Medicare fee-for-service (FFS) patients (2008-2010) demonstrate a median hospital-level RSCR of 3.5% (range 1.8% to 8.9%) after elective primary THA and/or TKA, suggesting room for improvement in clinical care.

The variation in complication rates across hospitals suggests there are considerable differences in the quality of care at the hospital level. Measuring and reporting risk-standardized complications rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and promote improvements in the quality of care received by patients and the outcomes they experience. The measure will also provide patients with information that could guide their choices regarding where they seek care for these elective procedures. Furthermore, the measure will increase transparency for consumers and has the potential to lower health care costs due to complications associated with these procedures.

1c.4. Citations for data demonstrating high priority provided in 1a.3

Bongartz T, Halligan CS, Osmon DR, et al. Incidence and risk factors of prosthetic joint infection after total hip or knee replacement in patients with rheumatoid arthritis. Arthritis Rheum. Dec 15 2008;59(12):1713-1720.

Bozic KJ, Rubash HE, Sculco TP, Berry DJ. An analysis of medicare payment policy for total joint arthroplasty. J Arthroplasty. Sep 2008;23(6 Suppl 1):133-138.

Browne J, Cook C, Hofmann A, Bolognesi M. Postoperative morbidity and mortality following total knee arthroplasty with computer navigation. Knee. Mar 2010;17(2):152-156.

Cram P, Vaughan-Sarrazin MS, Wolf B, Katz JN, Rosenthal GE. A comparison of total hip and knee replacement in specialty and general hospitals. J Bone Joint Surg Am. Aug 2007;89(8):1675-1684.

Huddleston JI, Maloney WJ, Wang Y, Verzier N, Hunt DR, Herndon JH. Adverse Events After Total Knee Arthroplasty: A National Medicare Study. The Journal of Arthroplasty. 2009;24(6, Supplement 1):95-100.

Khatod M, Inacio M, Paxton EW, et al. Knee replacement: epidemiology, outcomes, and trends in Southern California: 17,080 replacements from 1995 through 2004. Acta Orthop. Dec 2008;79(6):812-819.

Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am. Apr 2007a;89(4):780-785.

Kurtz SM, Ong KL, Schmier J, et al. Future clinical and economic impact of revision total hip and knee arthroplasty. J Bone Joint Surg Am. Oct 2007b;89 Suppl 3:144-151.

Kurtz S, Ong K, Lau E, Bozic K, Berry D, Parvizi J. Prosthetic joint infection risk after TKA in the Medicare population. Clin Orthop Relat Res. 2010;468:5.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. J Bone Joint Surg Am. Jan 2003;85-A(1):27-32.

Ong KL, Mowat FS, Chan N, Lau E, Halpern MT, Kurtz SM. Economic burden of revision hip and knee arthroplasty in Medicare enrollees. Clin Orthop Relat Res. May 2006;446:22-28.

Solomon DH, Chibnik LB, Losina E, et al. Development of a preliminary index that predicts adverse events after total knee replacement. Arthritis & Rheumatism. 2006;54(5):1536-1542.

Soohoo NF, Farng E, Lieberman JR, Chambers L, Zingmond DS. Factors That Predict Short-term Complication Rates After Total Hip Arthroplasty. Clin Orthop Relat Res. Sep 2010;468(9):2363-2371.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

N/A. This measure is not a PRO-PM.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Musculoskeletal : Joint Surgery, Musculoskeletal : Osteoporosis, Musculoskeletal : Rheumatoid Arthritis, Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Overuse, Safety, Safety : Complications, Safety : Venous Thromboembolism

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890569445&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DDelv21f_AUS_Procedure_Specific+Complications.pdf&blobcol=urldata&blobtable=MungoBlobs

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_1550_HipKnee_Complication_Data_Dictionary_v1.0.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Updates by Year

2016
No updates were made to the specifications of the THA/TKA complication measure for 2016 public reporting.

2015
1. Updated measure specifications to exclude patients from the cohort (denominator) without at least 90 days post-discharge enrollment in FFS Medicare.
Rationale: Removing index admissions for patients who withdrew from the Medicare FFS program within 90 days after discharge improves the accuracy of the measure by removing patients for whom there is no available outcome data and makes the measure consistent with the methodologies used in the THA/TKA 30-day readmission measure and other publicly reported condition-specific readmission measures for AMI, HF, pneumonia, COPD, and stroke admissions.

2014
1. Updated measure specifications to not include all patients with a secondary diagnosis of fracture during index admission in the measure cohort.
Rationale: These procedures are presumably not elective THA/TKA procedures and the cohort aims to include only elective THA/TKA procedures.

2. Updated measure specifications to exclude complications coded as POA during index admission from measure outcome.
Rationale: These complications are presumably not related to the index procedure and/or peri-operative care provided and the measure aims to assess quality of hospital care.

2013
1. Updated CC map
Rationale: Prior to 2014, the ICD-9-CM CC map was updated annually to capture all relevant comorbidities coded in patient administrative claims data.

2. Updated complication and fracture exclusion codes
Rationale: RTI identified new ICD-9-CM codes to add to the THA/TKA complication measure.
i. Updated ICD-9-CM codes defining the pneumonia, sepsis/septicemia, and pulmonary embolism complications to reflect changes to the ICD-9-CM coding (no change to the clinical meaning of the complications).
ii. Updated ICD-9-CM codes defining the femur, hip, or pelvic fracture exclusions to the measure cohort to reflect relevant new ICD-9-CM codes.

3. Changes from prior methodology report.
Rationale: Two tables were corrected from the original methodology report and the combined dataset was shortened from 36 to 33 months due to the timing of public reporting and the longer period of outcome assessment required to adequately capture complications up to 90 days following admission.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.*
The outcome for this measure is any complication occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. Complications are counted in the measure only if they occur during the index hospital admission or during a readmission. The complication outcome is a dichotomous (yes/no) outcome. If a patient experiences one or more of these complications in the applicable time period, the complication outcome for that patient is counted in the measure as a "yes".

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator Time Window: The specific time frame for the complication varies (depending on the complication) from (and including) the index admission to 90 days post date of the index admission (see details below in S.6 Numerator Details).

Denominator Time Window: The time window can be specified from one to three years. The measure is currently publicly-reported using three years of data.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population*

with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The composite complication is a dichotomous outcome (yes for any complication(s); no for no complications). Therefore, if a patient experiences one or more complications, the outcome variable will get coded as a "yes". Complications are counted in the measure only if they occur during the index hospital admission (and are not present on admission) or during a readmission.

The complications captured in the numerator are identified during the index admission OR associated with a readmission up to 90 days post-date of index admission, depending on the complication. The follow-up period for complications from date of index admission is as follows:

The follow-up period for AMI, pneumonia, and sepsis/septicemia/shock is seven days from the date of index admission because these conditions are more likely to be attributable to the procedure if they occur within the first week after the procedure. Additionally, analyses indicated a sharp decrease in the rate of these complications after seven days.

Death, surgical site bleeding, and pulmonary embolism are followed for 30 days following admission because clinical experts agree these complications are still likely attributable to the hospital performing the procedure during this period and rates for these complications remained elevated until roughly 30 days post admission.

The measure follow-up period is 90 days after admission for mechanical complications and periprosthetic joint infection/wound infection. Experts agree that mechanical complications and periprosthetic joint infection/wound infections due to the index THA/TKA occur up to 90 days following THA/TKA.

The measure counts all complications occurring during the index admission regardless of when they occur. For example, if a patient experiences an AMI on day 10 of the index admission, the measure will count the AMI as a complication, although the specified follow-up period for AMI is seven days. Clinical experts agree with this approach, as such complications likely represent the quality of care provided during the index admission.

As of 2014 reporting, the measure does not count complications in the complications outcome that are coded as POA during the index admission; this prevents identifying a condition as a complication of care if it was present on admission for the THA/TKA procedure.
For full list of ICD-9 and ICD-10 codes defining complications, see the Data Dictionary attached in field S.2b., sheet "Complication Codes ICD9-ICD10".

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
The target population for the publically reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures.

Additional details are provided in S.9 Denominator Details.

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
To be included in the measure cohort used in public reporting, patients must meet the following additional

inclusion criteria:

1. Enrolled in Medicare fee-for-service (FFS) Part A and Part B for the 12 months prior to the date of admission; and enrolled in Part A during the index admission;
2. Aged 65 or older
3. Having a qualifying elective primary THA/TKA procedure; elective primary THA/TKA procedures are defined as those procedures without any of the following:
• Femur, hip, or pelvic fractures coded in the principal or secondary discharge diagnosis field of the index admission
• Partial hip arthroplasty (PHA) procedures (with a concurrent THA/TKA); partial knee arthroplasty procedures are not distinguished by ICD9 codes and are currently captured by the THA/TKA measure
• Revision procedures with a concurrent THA/TKA
• Resurfacing procedures with a concurrent THA/TKA
• Mechanical complication coded in the principal discharge
• Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field
• Removal of implanted devises/prostheses
• Transfer status from another acute care facility for the THA/TKA

Patients are eligible for inclusion in the denominator if they had an elective primary THA and/or a TKA AND had continuous enrollment in Part A and Part B Medicare fee-for-service (FFS) 12 months prior to the date of index admission.

This measure can also be used for an all-payer population aged 18 years and older. We have explicitly tested the measure in both patients aged 18+ years and those aged 65+ years (see Section 2b4.11 of the Testing Attachment for details, 2b4.11).

International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes used to define the cohort for each measure are:
ICD-9-CM codes used to define a THA or TKA:
81.51 Total Hip Replacement
81.54 Total Knee Replacement

ICD-10 Codes that define a THA or TKA:
0SR90J9 Replacement of Right Hip Joint with Synthetic Substitute, Cemented, Open Approach
0SR90JA Replacement of Right Hip Joint with Synthetic Substitute, Uncemented, Open Approach
0SR90JZ Replacement of Right Hip Joint with Synthetic Substitute, Open Approach
0SRB0J9 Replacement of Left Hip Joint with Synthetic Substitute, Cemented, Open Approach
0SRB0JA Replacement of Left Hip Joint with Synthetic Substitute, Uncemented, Open Approach
0SRB0JZReplacement of Left Hip Joint with Synthetic Substitute, Open Approach
0SRC07Z Replacement of Right Knee Joint with Autologous Tissue Substitute, Open Approach
0SRC0JZReplacement of Right Knee Joint with Synthetic Substitute, Open Approach
0SRC0KZ Replacement of Right Knee Joint with Nonautologous Tissue Substitute, Open Approach
0SRD07Z Replacement of Left Knee Joint with Autologous Tissue Substitute, Open Approach
0SRD0JZ Replacement of Left Knee Joint with Synthetic Substitute, Open Approach
0SRD0KZReplacement of Left Knee Joint with Nonautologous Tissue Substitute, Open Approach
0SRT07Z Replacement of Right Knee Joint, Femoral Surface with Autologous Tissue Substitute, Open Approach
0SRT0JZ Replacement of Right Knee Joint, Femoral Surface with Synthetic Substitute, Open Approach
0SRT0KZ Replacement of Right Knee Joint, Femoral Surface with Nonautologous Tissue Substitute, Open Approach
0SRU07Z Replacement of Left Knee Joint, Femoral Surface with Autologous Tissue Substitute, Open Approach
0SRU0JZ Replacement of Left Knee Joint, Femoral Surface with Synthetic Substitute, Open Approach
0SRU0KZ Replacement of Left Knee Joint, Femoral Surface with Nonautologous Tissue Substitute, Open Approach

0SRV07Z Replacement of Right Knee Joint, Tibial Surface with Autologous Tissue Substitute, Open Approach
0SRV0JZ Replacement of Right Knee Joint, Tibial Surface with Synthetic Substitute, Open Approach
0SRV0KZ Replacement of Right Knee Joint, Tibial Surface with Nonautologous Tissue Substitute, Open Approach
0SRW07Z Replacement of Left Knee Joint, Tibial Surface with Autologous Tissue Substitute, Open Approach
0SRW0JZ Replacement of Left Knee Joint, Tibial Surface with Synthetic Substitute, Open Approach
0SRW0KZ Replacement of Left Knee Joint, Tibial Surface with Nonautologous Tissue Substitute, Open Approach

An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).

Elective primary THA/TKA procedures are defined as those procedures without any of the following:
1) Femur, hip, or pelvic fractures coded in principal or secondary discharge diagnosis fields of the index admission
2) Partial hip arthroplasty (PHA) procedures with a concurrent THA/TKA
3) Revision procedures with a concurrent THA/TKA
4) Resurfacing procedures with a concurrent THA/TKA
5) Mechanical complication coded in the principal discharge
6) Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field
7) Removal of implanted devises/prostheses
8) Transfer status from another acute care facility for the THA/TKA

For a full list of ICD-9 and ICD-10 codes defining the following see attached Data Dictionary, sheet "THA TKA Cohort Codes Part 2."

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare;
2. Who were discharged against medical advice (AMA); or,
3. Who had more than two THA/TKA procedure codes during the index hospitalization.

After applying these exclusion criteria, we randomly select one index admission for patients with multiple index admissions in a calendar year. We therefore exclude the other eligible index admissions in that year.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare
Rationale: The 90-day complication outcome cannot be assessed in this group since claims data are used to determine whether a complication of care occurred.

2. Who were discharged against medical advice (AMA); or,
Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge.

3. Who had more than two THA/TKA procedure codes during the index hospitalization
Rationale: Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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*

with at S.2b)
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Our approach to risk adjustment is tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al., 2006).

The measure employs a hierarchical logistic regression model to create a hospital-level RSCR. In brief, the approach simultaneously models data at the patient and hospital levels to account for the variance in patient outcomes within and between hospitals (Normand & Shahian, 2007). At the patient level, the model adjusts the log-odds of complications occurring within 90 days of the index admission using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, the approach models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of complication at the hospital, after accounting for patient risk. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Candidate and Final Risk-adjustment Variables: Candidate variables were patient-level risk-adjustors that were expected to be predictive of complication, based on empirical analysis, prior literature, and clinical judgment, including age and indicators of comorbidity and disease severity. For each patient, covariates are obtained from claims records extending 12 months prior to and including the index admission. For the measure currently implemented by CMS, these risk adjusters are identified using both inpatient and outpatient Medicare FFS claims data. However, in the all-payer hospital discharge database measure, the risk-adjustment variables can be obtained only from inpatient claims in the prior 12 months and the index admission.

The model adjusts for case-mix differences based on the clinical status of patients at the time of admission. We use condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes (Pope et al., 2000). A file that contains a list of the ICD-9-CM codes and their groupings into CCs is attached in data field S.2b (Data Dictionary or Code Table). In addition, only comorbidities that convey information about the patient at admission or in the 12 months prior, and not complications that arise during the course of the index hospitalization, are included in the risk adjustment. Hence, we do not risk adjust for CCs that may represent adverse events of care when they are only recorded in the index admission.

The final set of risk-adjustment variables is:

Demographics
Age-65 (years, continuous) for patients aged 65 or over cohorts; or Age (years, continuous) for patients aged 18 and over cohorts
Male (%)

THA/TKA Procedure
Index admissions with an elective THA procedure
Number of procedures (two vs. one)

Clinical Risk Factors

Other congenital deformity of hip (joint) (ICD-9 code 755.63)
Post traumatic osteoarthritis (ICD-9 codes 716.15, 716.16)
Morbid obesity (ICD-9 code 278.01)
Metastatic cancer or acute leukemia (CC 7)
Cancer (CC 8-12)
Respiratory/heart/digestive/urinary/other neoplasms (CC 11-13)
Diabetes mellitus (DM) or DM complications (CC 15-20, 119, 120)
Protein-calorie malnutrition (CC 21)
Bone/joint/muscle infections/necrosis (CC 37)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)
Osteoarthritis of hip or knee (CC 40)
Osteoporosis and other bone/cartilage disorders (CC 41)
Dementia or other specific brain disorders (CC 49-50)
Major psychiatric disorders (CC 54-56)
Hemiplegia, paraplegia, paralysis, function disability (CC 67-69, 100-102, 177-178)
Cardio-respiratory failure and shock (CC 79)
Coronary atherosclerosis or angina (CC 83-84)
Stroke (CC 95-96)
Vascular or circulatory disease (CC 104-106)
Chronic obstructive pulmonary disease (COPD) (CC 108)
Pneumonia (CC 111-113)
Pleural effusion/pneumothorax (CC 114)
Dialysis status (CC 130)
Renal failure (CC 131)
Decubitus ulcer or chronic skin ulcer (CC 148-149)
Trauma (CC 154-156, 158-161)
Vertebral fractures (CC 157)
Other injuries (CC 162)
Major complications of medical care and trauma (CC 164)

References:
Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

Pope G,Ellis R,Ash A, et al. Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment. Health Care Financing Review. 2000;21(3):26.

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The measure estimates hospital-level RSCRs following elective primary THA/TKA using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals (Normand and Shahian, 2007). At the patient level, it models the log-odds of a complication occurring within 90 days of the index admission using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, it models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a complication at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

The RSCR is calculated as the ratio of the number of "predicted" to the number of "expected" admissions with a complication at a given hospital, multiplied by the national observed complication rate. For each hospital, the numerator of the ratio is the number of complications within 90 days predicted on the basis of the hospital's performance with its observed case mix, and the denominator is the number of complications expected based on the nation's performance with that hospital's case mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows for a comparison of a particular hospital's performance given its case mix to an average hospital's performance with the same case mix. Thus, a lower ratio indicates lower-than-expected complication rates or better quality, and a higher ratio indicates higher-than-expected complication rates or worse quality.

The "predicted" number of admissions with a complication (the numerator) is calculated by using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of having an admission with a complication. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are log transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of admissions with a complication (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific effect. The results are log transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the years of data in that period.

This calculation transforms the ratio of predicted over expected into a rate that is compared to the national observed complication rate. The hierarchical logistic regression models are described fully in the original methodology report (Grosso et al., 2012).

References:
Grosso L, Curtis J, Geary L, et al. Hospital-level Risk-Standardized Complication Rate Following Elective Primary Total Hip Arthroplasty (THA) And/Or Total Knee Arthroplasty (TKA) Measure Methodology Report. 2012.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22(2): 206-226.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. This measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A. This measure is not based on a survey or patient-reported data.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing values are rare among variables used from claims data in this measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Other, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Data sources:

The currently publically reported measure is specified and has been tested using:
1. Medicare Part A inpatient and Part B outpatient claims: This data source contains claims data for FFS inpatient and outpatient services including: Medicare inpatient hospital care, outpatient hospital services, as well as inpatient and outpatient physician claims for the 12 months prior to an index admission.

2. Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission as well as vital status at discharge. These data have previously been shown to accurately reflect patient vital status (Fleming et al., 1992).

During original measure development we validated the administrative claims-based definition of THA/TKA complication (original model specification) against a medical record data.
3. Data abstracted from medical records from eight participating hospitals (approximately 96 records per hospital; 644 total records) for Medicare beneficiaries over the age of 65 years who had a qualifying THA/TKA procedure between January 1 2007 and December 31, 2008.

The measure was also specified and testing using an all-payer claims dataset although it is only publically reported using the data sources listed above
4. California Patient Discharge Data is a large, linked database of patient hospital admissions in the state of California. Using all-payer data from California, we performed analyses to determine whether the THA/TKA complication measure can be applied to all adult patients, including not only FFS Medicare patients aged 65 years or over, but also non-FFS Medicare patients aged 18-64 years at the time of admission.

Additional Data source used for analysis of the impact of SES variables on the measure's risk model. Note, the variables derived from these data are not included in the measure as specified
5. The American Community Survey (2009-2013): The American Community Survey data is collected annually and an aggregated 5-years data was used to calculate the AHRQ socioeconomic status (SES) composite index score.

Reference:
Fleming C., Fisher ES, Chang CH, Bubolz D, Malenda J. Studying outcomes and hospital utilization in the elderly: The advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Medical Care. 1992; 30(5): 377-91.

Suter LG, Parzynski CS, Grady JN, et al. 2014 Procedure Specific Complication Measure Updates and Specifications Report: Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) Risk-Standardized Complication Measure (Version 3.0). 2014

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A. This measure is not a composite performance measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_1550_HipKnee_Complication_NQF_Testing_Attachment_v1.1.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 1550
Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
Date of Submission: 5/31/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about

how to present all the testing information in one form.

- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

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

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

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

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

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

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

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Census Data/American Community Survey	☒ other: Census Data/American Community Survey

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

The datasets used for testing included Medicare Parts A and B claims, as well as the Medicare Enrollment Database (EDB). Additionally, census data were used to assess socioeconomic factors and race (dual eligibility and African American race variables obtained through enrollment data; Agency for Healthcare Research and Quality [AHRQ] socioeconomic status [SES] index score obtained through census data). Data abstracted from hospital medical records were used to validate the claims-based assessment of the complication outcome. The dataset used varies by testing type; see Section 1.7 for details.

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

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

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

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item***	**Measure Tested at Level of:**

S.26)	
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 47T	☐ other: 47T

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

For this measure, hospitals are the measured entities. All non-federal, acute inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

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

The number of admissions/patients varies by testing type: see Section 1.7 for details

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

The datasets, dates, number of measured entities and number of admissions used in each type of testing are as follows:

For reliability testing (Section 2a2)

For reliability testing, we randomly split Dataset 1 into two samples. The reliability of the model was tested by randomly selecting 50% of the Medicare patients aged 65 years and over in the most recent three-year cohort and developing a risk-adjusted model for this group. We then developed a second model for the remaining 50% of patients and compared the two. In each year of measure reevaluation, we also re-fit the model and compared the frequencies and model coefficients of risk variables (condition categories for patient comorbidities) and model fit across 3 years (**Dataset 1** below).

Dataset 1 (2015 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims, and Medicare Enrollment Database (to assess enrollment in FFS)
Dates of Data: April 1, 2011 – March 30, 2014
Number of Index Admissions: 892,455
Patient Descriptive Characteristics: average age= 74.8, %male= 36.6

Number of Measured Entities: 3,507 hospitals

For validity testing (Section 2b2)

No empirical testing was done to assess measure score validity. However, data from hospital medical records were used to assess the validity of the complication outcome.

Dataset 2 (medical record validation of the complication outcome assessment): Hospital medical record data were compared with data from Medicare claims to assess agreement in the identification of complications following THA/TKA procedures
Dates of Data: January 1, 2007 – December 31, 2008
Number of Admissions: N=644
Number of Measured Entities: 8 acute-care hospitals

For testing of measure exclusions (Section 2b3)

Dataset 1 (2015 public reporting cohort)

For testing of measure risk adjustment (Section 2b4)

Dataset 1 (2015 public reporting cohort)

Dataset 3 (development dataset): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims
Dates of Data: 2008
Number of Admissions: N=145,206 (first half of split sample); N=145,123 (second half of split sample)
Patient Descriptive Characteristics: average age=75.2 (first half of split sample), %male=35.8 (first half of split sample); average age=75.2 (second half of split sample), %male=35.6 (second half of split sample)
Number of Measured Entities: 3,221 hospitals (first half of split sample); 3,223 hospitals (second half of split samples)

To create the model development sample (**Dataset 3**), we applied the inclusion and exclusion criteria to all 2008 admissions. We randomly selected half of all THA/TKA admissions in 2008 that met the inclusion and exclusion criteria to create a model development sample. We used the remaining admissions as our model validation sample.

For Sub-section 2b4.11. Optional Additional Testing for Risk Adjustment

Dataset 4 (all payer dataset): California Patient Discharge Data in addition to CMS Medicare FFS inpatient claims data used to test the measure in patients 18 years and older
Dates of Data: January 1, 2006 – December 31, 2006
Number of Index Admissions: 59,828 (all patients 18 years and older) [mean age=67.2, %male=39.8]

For testing to identify meaningful differences in performance (Section 2b5)

Dataset 1 (2015 public reporting cohort)

For testing of sociodemographic factors in risk models (Section 2b4.4b)

Dataset 1 (2015 public reporting cohort); **Dataset 5** (The American Community Survey [ACS]): The American Community Survey, 2009-2013

We examined disparities in performance according to the proportion of patients in each hospital who were of African-American race and the proportion who were dual eligible for both Medicare and Medicaid insurances. We also used the AHRQ SES index score to study the association between performance measures and socioeconomic status (SES).

Data Elements

• African-American race and dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data are obtained from CMS enrollment data (**Dataset 1**)
• Validated AHRQ SES index score is a composite of 7 different variables found in the census data (**Dataset 5**)

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

SDS incorporates socioeconomic variables as well as race into a more concise term. However, given the fact that socioeconomic risk factors are distinct from race and should be interpreted differently, we have decided to keep "SES" and "race" as separate terms.

We selected SES and race variables to analyze after reviewing the literature and examining available national data sources. There is a large body of literature linking various SES factors and African-American race to worse health status, higher mortality over a lifetime, and hospital outcomes such as readmission (Adler and Newman, 2002; Blum et al., 2014; Eapen et al. 2015; Gilman et al., 2014; Hu et al., 2014; Joynt and Jha, 2013; Mackenbach et al., 2000; Tonne et al., 2005; van Oeffelen et al., 2012). Income, education, and occupational level are the most commonly examined variables. Although literature directly examining how different SES factors or race might influence the likelihood of older, insured, Medicare patients experiencing complications following admission for hip/knee surgery is more limited, studies have indicated an association between SES and increased risk of postoperative hip/knee surgery complications (Browne, Novicoff, and D'Apuzzo 2014; Ong et al., 2009), while others have found similar rates of complications for hospitals caring for higher and lower proportions of low SES patients or African-American patients (Bozic et al., 2014). In addition, studies have also suggested other disparities related to hip/knee surgery, including significant differences in the rate of total hip replacement surgery received by African-American and white patients (Ibrahim, 2010; Mahomed, 2003). The causal pathways for SES and race variable selection are described below in Section 2b4.3.

The SES and race variables used for analysis were:

- Dual eligible status (**Dataset 1**)
- African-American race (**Dataset 1**)
- AHRQ-validated SES index score (summarizing the information from the following variables: percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12thgrade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room) (**Dataset 5**)

In selecting variables, our intent was to be responsive to the NQF guidelines for measure developers in the context of the SDS Trial Period. Our approach has been to examine all patient-level indicators of both SES and race that are reliably available for all Medicare beneficiaries, are linkable to claims data, and have established validity.

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

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

Finally, we selected the AHRQ-validated SES index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas (Bonito et al., 2008). Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. In this submission, we present analyses using the census block level, the most granular level possible using ACS data. A census block group is a geographical unit used by the US Census Bureau which is between the census tract and the census block. It is the smallest geographical unit for which the bureau publishes sample data. The target size for block groups is 1,500 and they typically have a population of 600 to 3,000 people. We used 2009-2013 ACS data and mapped patients' 9-digit ZIP codes via vendor software to the AHRQ SES Index at the census block group level. Given the variation in cost of living across the country, the median income and median property value components of the AHRQ SES Index were adjusted by regional price parity values published by the Bureau of Economic Analysis (BEA). This provides a better marker of low SES neighborhoods in high expense geographic areas. We then calculated an AHRQ SES Index score for census block groups that can be linked to 9-digit ZIP codes. In the THA/TKA measure cohort, we were able to assign an AHRQ SES Index score to 99.6% of patient admissions. 88.6% of patient admissions had calculated AHRQ SES Index scores linked to their 9-digit ZIP codes. 11.0% of patient admissions had only valid 5-digit ZIP codes; we utilized the data for the median 9-digit ZIP code within that 5-digit ZIP code.

References:

Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. Health affairs (Project Hope). 2002; 21(2):60-76.

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation. Cardiovascular quality and outcomes. May 2014; 7(3):391-397.

Bozic KJ, Grosso LM, Lin Z, et al. Variation in Hospital-Level Risk-Standardized Complication Rates Following Elective Primary Total Hip and Knee Arthroplasty. The Journal of Bone & Joint Surgery .2014;96:640-647

Bonito A, Bann C, Eicheldinger C, Carpenter L. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final Report, Sub-Task. 2008;2.

Browne JA, Novicoff WM, D'Apuzzo MR. Medicaid payer status is associated with in-hospital morbidity and resource utilization following primary total joint arthroplasty. The Journal of bone and joint surgery. American volume. 2014;96(21):e180.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, HernandezAF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

Ibrahim SA. Racial variations in the utilization of knee and hip joint replacement: an introduction and review of the most recent literature. Current orthopaedic practice. 2010;21:126-131

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Mackenbach JP, Cavelaars AE, Kunst AE, Groenhof F. Socioeconomic inequalities in cardiovascular disease mortality; an international study. European heart journal. 2000; 21(14):1141-1151.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. The Journal of bone and joint surgery American volume. 2003;85-a:27-32

Ong KL, Kurtz SM, Lau E, Bozic KJ, Berry DJ, Parvizi J. Prosthetic joint infection risk after total hip arthroplasty in the Medicare population. The Journal of arthroplasty. 2009;24(6 Suppl):105-109.

Tonne C, Schwartz J, Mittleman M, Melly S, Suh H, Goldberg R. Long-term survival after acute myocardial infarction is lower in more deprived neighborhoods. Circulation. Jun 14 2005; 111(23):3063-3070.

van Oeffelen AA, Agyemang C, Bots ML, et al. The relation between socioeconomic status and short-term mortality after acute myocardial infarction persists in the elderly: results from a nationwide study. European journal of epidemiology. Aug 2012; 27(8):605-613.

2a2. RELIABILITY TESTING

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

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

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

Data Element Reliability

In constructing the measure, we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across hospitals or providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard. For example, "discharge disposition" is a variable in Medicare claims data that is not thought to be a reliable variable for identifying a transfer between two acute care facilities. Thus, we derive a variable using admission and discharge dates as a surrogate for "discharge disposition" to identify hospital admissions involving transfers. This allows us to identify these admissions using variables in the claims data which have greater reliability than the "discharge disposition" variable.

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

Finally, we assess the reliability of the data elements by comparing model variable frequencies and odds ratios from logistic regression models across the most recent three years of data (**Dataset 1**, see section 1.7).

Measure Score Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability was to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. That is, we took a "test-retest" approach in which hospital performance was measured once using a random subset of patients, then measured again using a second random subset exclusive of the first. Finally, we compared the agreement between the two resulting performance measures across hospitals (Rousson et al., 2002).

For test-retest reliability, we combined index admissions from successive measurement periods into one dataset, randomly sampled half of patients within each hospital, calculated the measure for each

hospital, and repeated the calculation using the second half of patients. Thus, each hospital was measured twice, but each measurement was made using an entirely distinct set of patients. To the extent that the calculated measures of these two samples agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used the **Dataset 1** split sample and calculated the RSCR for each hospital for each sample. The agreement of the two RSCRs was quantified for hospitals using the ICC (2,1) as defined by Shrout and Fleiss (1979).

Using two independent samples provides a stringent estimate of the measure's reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ´signal´, a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman, 1910; Brown, 1910). We used this to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

References:

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002;21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

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

Data Element Reliability Results (**Dataset 1**)

The frequency of some model variables increased while others decreased between 2011 and 2014, which may reflect an increased or decreased rate of specific comorbidities in the Medicare FFS population. For example, there was a notable decrease (≥ 2%) in the frequency of coronary atherosclerosis or angina (CC 83-84) (29.0% to 26.7%). Examination of the odds ratios for each risk variable in the model shows that, overall, the odds ratios for individual risk variables remained relatively constant across the three years.

For the model variable frequencies, see the *2015 Measure Updates and Specifications Report* (Dorsey et al., 2016) attached to this submission.

Measure Score Reliability Results (Dataset 1)

There were 892,455 admissions in the 3-year split sample (from **Dataset 1**), with 445,352 index admissions from 2,826 hospitals in one sample and 447,103 admissions from 2,851 hospitals in the other randomly selected sample. The agreement between the two RSCRs for each hospital, the ICC, was 0.45, which according to the conventional interpretation is "moderate" (Landis & Koch, 1977).

Note that we limited this analysis to hospitals with 12 or more cases in each split sample.

The ICC is based on a split sample of three years of data, resulting in a volume of patients in each sample equivalent to only 1.5 years of data, whereas the measure is reported with the full three years of data.

References:

Dorsey K, Grady J, Suter LG, et al. 2016 Procedure-Specific Measure Updates and Specifications Report Hospital-Level Risk-Standardized Complication Measure: Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Version 5.0). 2016. https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890569445&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DDelv21f_AUS_Procedure_Specific+Complications.pdf&blobcol=urldata&blobtable=MungoBlobs. Accessed May 16, 2016.

Landis J, Koch G. The measurement of observer agreement for categorical data, Biometrics 1977;33:159-174.

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

The stability of the risk factor odds ratios over time suggests that the underlying data elements are reliable. Additionally, the ICC score demonstrates moderate agreement across samples using a conservative approach to assessment.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

- ☒ **Empirical validity testing**
- ☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Measure validity is demonstrated through prior validity testing or claims-based risk models done on our claims-based measures, through use of established measure development guidelines, and

by systematic assessment of measure face validity by a technical expert panel (TEP) of national experts and stakeholder organizations.

Validity of Other Claims-Based Measures:

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission measures) with models that used chart-abstracted data for risk adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical record data for risk adjustment for heart failure patients (National Heart Failure data) (Krumholz, Wang, et al. 2006; Keenan et al. 2008), AMI patients (Cooperative Cardiovascular Project data) (Krumholz, Wang, et al. 2006), and pneumonia patients (National Pneumonia Project dataset) (Bratzler et al. 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (elective primary THA/TKA and implantable cardioverter defibrillator). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical record data.

Validity Indicated by Established Measure Development Guidelines:

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, “Standards for Statistical Models Used for Public Reporting of Health Outcomes” (Krumholz, Brindis, et al. 2006).

Validity as Assessed by External Groups:

Throughout measure development, we obtained expert and stakeholder input via three mechanisms: regular discussions with an advisory working group, a national TEP, and a 15-day public comment period in order to increase transparency and to gain broader input into the measure.

We assembled the working group and held regular meetings throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians and other professionals with expertise in biostatistics, measure methodology, and quality improvement. Working group meetings addressed key issues related to measure development, including weighing the pros and cons of and finalizing key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader TEP.

In addition to the working group, and in alignment with the CMS MMS, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in

relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including physicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS website. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

Finally, NQF previously endorsed this measure in 2012, demonstrating additional external groups' endorsement of the measure's validity.

Face Validity as Determined by TEP:

One means of confirming the validity of this measure was face validity assessed by our TEP.

List of TEP Members

1. Mark L. Francis, MD
Professor of Medicine and Biomedical Sciences, Chief, Division of Rheumatology, Department of Internal Medicine, Texas Tech University Health Sciences Center

2. Cynthia Jacelon, PhD, RN, CRRN
Associate Professor, School of Nursing, University of Massachusetts; Association of Rehabilitation Nurses

3. Norman Johanson, MD
Chairman, Orthopedic Surgery, Drexel University College of Medicine

4. C. Kent Kwoh, MD
Professor of Medicine, Associate Chief and Director of Clinical Research, Division of Rheumatology and Clinical Immunology University of Pittsburgh

5. Courtland G. Lewis, MD
American Association of Orthopaedic Surgeons

6. Jay Lieberman, MD
Professor and Chairman, Department of Orthopedic Surgery, University of Connecticut Health Center; Director, New England Musculoskeletal Institute

7 Peter Lindenauer, MD, M.Sc.
Hospitalist and Health Services Researcher, Baystate Medical Center; Professor of Medicine, Tufts University

8. Russell Robbins, MD, MBA
Principal, Mercer's Total Health Management

9. Barbara Schaffer
THA Patient

10. Nelson SooHoo, MD, MPH
Professor, University of California at Los Angeles

11. Steven H. Stern, MD
Vice President, Cardiology & Orthopedics/ Neuroscience, United Healthcare

12. Richard E. White, Jr., MD
American Association of Hip and Knee Surgeons

Empirical Validation of Claims-Based Definition of Complications

During original measure development we validated the administrative claims-based definition of THA/TKA complication (original model specification) against medical record data (**Dataset 2**). The primary goal of this validation study was to determine the overall agreement between patients identified as having a complication (or no complication) using claims data compared with those who had a complication (or no complication) documented in the medical record. We conducted a secondary analysis of agreement of individual, specific complications to identify opportunities for measure improvement.

A senior statistician conducted a detailed analysis of each abstracted patient record and compared the findings to the patient results found in the claims data. If any disagreement between the medical record abstraction and the claims data was found, the disagreement was documented and explored in further detail. In some instances, we requested that the medical record be re-abstracted in order to confirm the disagreement and/or to obtain more clinical information. Our clinical team also reviewed some medical records to further determine the nature of disagreement.

To determine overall measure agreement, we calculated the percentage of patients for whom both the claims and medical record identified at least one complication or neither identified a complication. For each case where there was a disagreement between the medical record and claims-based measure, we verified and characterized each disagreement. We then conducted a detailed review of all disagreements between the specific complications documented (or not documented) in the claims data and the medical records, even if such disagreements did not result in overall measure disagreement. We then calculated the percentage of patients where the exact complication(s) coded in claims was also documented in the medical record and vice versa (referred to throughout as “one-to-one agreement”).

References:

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.

Shahian DM, He X, O'Brien S, et al. Development of a Clinical Registry-Based 30-Day Readmission Measure for Coronary Artery Bypass Grafting Surgery. Circulation 2014; DOI: 0.1161/CIRCULATIONAHA.113.007541. Published online before print June 10, 2014

ICD-9 to ICD-10 Conversion

Statement of Intent

[X] Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.
[ ] Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.
[ ] The intent of the measure has changed.

Process of Conversion

ICD-10 codes were initially identified using Generalized Equivalency Mapping (GEM) software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes currently in use for this measure. Each year we reexamine the codes using the latest version of the GEM software. We completed this examination most recently in 2015. An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Validity as Assessed by External Groups

Although there are no empiric results for measure score validity, the commenters provided valuable insights and recommendations for measure improvement. Comments regarding the complication measure were primarily in reference to risk adjustment for SES and comorbidities, complications included/excluded in the measure, and the limitations of claims data in capturing comorbidities and defining complications. Overall, responders supported the complication measure. In addition, the TEP provided input about the risk model that was incorporated in the model to strengthen the measure.

Validation of Claims-Based Definition of Complications

Overall measure agreement was 93% (598/644 patients). More specifically, there were 598 patients who either had a complication coded in the claims and a complication was also documented in the medical record or who had no complication documented in both claims and medical record data. When we examined overall agreement in patients with and without complications, initial agreement was 86% for patients with a complication compared with 99%

for patients without a complication. We proposed some minor changes to the measure on the basis of this validation study. Specifically, we determined that ICD-9 code 998.59, "Other postoperative infection," was not sufficiently specific to sepsis, and the measure identified cases of sepsis that were not documented in the medical record. Therefore, we recommended removal of this code from the measure specifications. Secondly, we recommended combining wound infection and periprosthetic joint infection as a single complication in the measure specifications because these complications can be clinically difficult to differentiate. After the proposed measure changes were implemented, measure agreement between claims data and the medical record will increase to 99% (635/644 patients).

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Validity as Assessed by External Groups

The TEP's feedback on the measure demonstrated their agreement with the overall face validity of the measure as specified.

Validation of Claims-Based Definition of Complications

The administrative claims-based and medical record data showed a high level of agreement in how they identified complications in the validity testing that was performed. There was overall measure agreement between the claims data and the medical record on the measure outcome in 99% of the cases after improving the claims-based definition of complication.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions and to ensure accurate calculation of the measure. To ascertain the impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Dataset 1**). These exclusions are consistent with similar NQF-endorsed outcome measures. For more details see the attached specifications report.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

In **Dataset 1** (2015 public reporting cohort):

Exclusion	N	%	Distribution across hospitals (N=2,826): Minimum, 25th percentile, 50th percentile, 75th percentile, maximum
1. Discharged against medical advice (AMA)	112	0.01	(0.0, 0.0, 0.0, 0.0, 0.3)
2. Without at least 90 days post-discharge enrollment in FFS Medicare for index admissions	1,397	0.15	(0.0, 0.0, 0.0, 0.0, 8.6)
3. Admitted for the index procedure and subsequently transferred to another acute care facility	1,892	0.20	(0.0, 0.0, 0.0, 0.2, 5.5)
4. Had more than two THA/TKA procedure codes during the index hospitalization	1	<0.01	(0.0, 0.0, 0.0, 0.0, 0.2)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exclusion 1 (patients who are discharged AMA) accounts for 0.01% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to adequately deliver full care. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 2 (patients without at least 90 days of post-discharge enrollment in FFS Medicare for index admissions) accounts for 0.15% of all index admissions excluded from the initial cohort. This exclusion is needed because the 90-day complication outcome cannot be assessed in this group since claims data are used to determine whether a patient experienced complications. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 3 (patients who are transferred to another acute care facility) accounts for 0.20% of all index procedures excluded from the initial index cohort. This exclusion is intended to remove admissions from the cohort for patients transferred in to the index hospital, as they likely do not represent elective THA/TKA procedures.

Exclusion 4 (patients with more than two THA/TKA procedure codes during the index hospitalization) accounts for <0.01% of all index procedures excluded from the initial index cohort. Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

I***f not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 33 risk factors**

☐ **Stratification by 47T risk categories**

☐ **Other,** 47T

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Our approach to risk adjustment was tailored to, and appropriate for, a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al. 2006).

The measure employs a hierarchical logistic regression model (a form of hierarchical generalized linear model [HGLM]) to create a hospital-level 90-day RSCR. This approach to modeling appropriately accounts for the structure of the data (patients clustered within hospitals), the underlying risk due to patients' comorbidities, and sample size at a given hospital when estimating hospital complication rates. In brief, the approach simultaneously models two levels (patient and hospital) to account for the variance in patient outcomes within and between hospitals (Normand and Shahian et al. 2007). At the patient level, each model adjusts the log odds of complications within 90 days of admission for THA/TKA procedure for age, sex, selected clinical covariates, and a hospital-specific intercept. The second level models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept, or hospital-specific effect, represents the hospital contribution to the risk of complications, after accounting for patient risk and sample size, and can be inferred as a measure of quality. The hospital-specific intercepts are given a distribution in order to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Clinical Factors

Candidate and Final Risk-adjustment Variables

The original measure was developed using Medicare FFS claims data. Candidate variables were patient-level risk adjustors that were expected to be predictive of complications, based on empirical analysis, prior literature, and clinical judgment, including demographic factors (age, sex) and indicators of comorbidity and disease severity. For each patient, covariates were obtained from Medicare claims extending 12 months prior to and including the index admission. The model adjusted for case differences based on the clinical status of the patient at the time of admission. We used condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes. We did not risk adjust for CCs that were possible adverse events of care and that were only recorded in the index admission. In addition, only comorbidities that conveyed information about the patient at that time or in the 12 months prior, and not complications that arose during the course of the admission were included in the risk adjustment.

The final set of risk-adjustment variables is:

Demographic

- Age-65 (years above 65, continuous)
- Male

THA/TKA Procedure

- Index admissions with an elective THA procedure
- Number of procedures performed

Clinical Risk Factors

- Other congenital deformity of hip (joint) (ICD-9 code 755.63)
- Post traumatic osteoarthritis (ICD-9 codes 716.15, 716.16)
- Morbid obesity (ICD-9 code 278.01)
- Metastatic cancer or acute leukemia (CC 7)
- Cancer (CC 8-10)
- Respiratory/heart/digestive/urinary/other neoplasms (CC 11-13)
- Diabetes mellitus (DM) or DM complications (CC 15-20, 119-120)
- Protein-calorie malnutrition (CC 21)
- Bone/joint/muscle infections/necrosis (CC 37)
- Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)
- Osteoarthritis of hip or knee (CC 40)
- Osteoporosis and other bone/cartilage disorders (CC 41)
- Dementia or other specified brain disorders (CC 49-50)
- Major psychiatric disorders (CC 54-56)
- Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)
- Cardio-respiratory failure or shock (CC 79)
- Coronary atherosclerosis or angina (CC 83-84)
- Stroke (CC 95-96)
- Vascular or circulatory disease (CC 104-106)
- Chronic obstructive pulmonary disease (COPD) (CC 108)
- Pneumonia (CC 111-113)
- Pleural effusion/pneumothorax (CC 114)
- Dialysis status (CC 130)

• Renal failure (CC 131)
• Decubitus ulcer or chronic skin ulcer (CC 148-149)
• Trauma (CC 154-156, 158-161)
• Vertebral fractures (CC 157)
• Other injures (CC 162)
• Major complications of medical care and trauma (CC 164)

SES Factors and Race

We selected variables representing SES factors and race for examination based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below we describe the pathways by which SES and race may influence 90-day complication rates following primary elective THA/TKA procedures.

Our conceptualization of the pathways by which patient SES or race affects 90-day complications is informed by the literature.

Literature Review of SES and Race Variables and THA/TKA Complications

To examine the relationship between SES and race variables and hospital 90-day RSCR following a THA/TKA procedure, a literature search was performed with the following exclusion criteria: international studies, articles published more than 10 years ago, articles without primary data, articles using Veterans Affairs databases as the primary data source, and articles not explicitly focused on SES or race and hip/knee surgery complications. Thirty-six studies were initially reviewed, and 31 studies were excluded from full-text review based on the above criteria. While several studies indicated that SES variables were associated with increased risk of postoperative hip/knee surgery complications (Browne, Novicoff, and D'Apuzzo 2014; Cram et al., 2007; Katz, Bierbaum, and Losina 2008; Ong et al., 2009), others have found similar rates of complications for hospitals caring for higher and lower proportions of low SES patients or African-American patients (Bozic et al., 2014). In addition to the literature focused on complications following hip/knee surgery, other studies have also found significant differences in the rate of THA received by African-American and white patients, indicating that patient and surgeon behavior may also contribute to disparities based on racial factors (Ibrahim, 2010; Mahomed et al., 2003).

Causal Pathways for SES and Race Variable Selection

Although some recent literature evaluates the relationship between patient SES or race and postoperative complications (including mortality) or outcomes such as readmission across conditions and procedures, few studies directly address causal pathways or examine the role of the hospital in these pathways. Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with complications. The SES factors that have been examined in complication literature, which spans across conditions and procedures, can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables. Patient-level variables describe characteristics of individual patients, and range from the self-reported or documented race or ethnicity of the patient to the patient's income or education level (Alter et al., 2014; Cram et al., 2007; Taksler et al., 2012). Neighborhood/community-level variables use information from sources such as the ACS as

either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the AHRQ-validated SES index score (Blum et al., 2014). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital.

The conceptual relationship, or potential causal pathways by which these possible SES risk factors influence the risk of complications following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider.

1. **Relationship of SES factors or race to health at admission**. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their hospitalization or procedure with a greater severity of underlying illness. These SES risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job, lack of childcare), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment**.**

In addition to SES risk factors, studies have shown that worse health status is more prevalent among African-American patients compared with white patients. The association between race and worse health is in part mediated by the association between race and SES risk factors such as poverty or disparate access to care associated with poverty or neighborhood. The association is also mediated through bias in healthcare as well as other facets of society.

2. **Use of low-quality hospitals**. Patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high quality facilities because such facilities are less likely to be found in geographic areas with large populations of poor patients; thus patients with low income are more likely to be seen in lower quality hospitals, which can contribute to increased risk of complications following hospitalization (Jha et al., 2011; Reames et al., 2014). Similarly, African-American patients have been shown to have less access to high quality facilities compared with white patients (Skinner et al., 2005).

3. **Differential care within a hospital**. The third major pathway by which SES factors or race may contribute to complication risk is that patients may not receive equivalent care within a facility. For example, African-American patients have been shown to experience differential, lower quality, or discriminatory care within a given facility (Trivedi et al., 2014). Alternatively, patients with SES risk factors such as lower education may require differentiated care – e.g. provision of lower literacy information – that they do not receive.

4. **Influence of SES on complication risk outside of hospital quality and health status**. Some SES risk factors, such as income or wealth, may affect the likelihood of complications without directly affecting health status at admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-surgery due to competing economic priorities or a lack of access to care outside of the hospital.

These proposed pathways are complex to distinguish analytically. They also have different implications on the decision to risk adjust or not. We, therefore, first assessed if there was evidence of a meaningful effect on the risk model to warrant efforts to distinguish among these pathways. Based on this model and the considerations outlined in Section 1.8, the following SES and race variables were considered:

- Dual-eligible status,
- African-American race, and
- AHRQ SES Index

We assessed the relationship between the SES variables and race with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

One concern with including SES or race factors in a model is that their effect may be at either the patient or the hospital level. For example, low SES may increase the risk of complications because patients of low SES have an individual higher risk (patient-level effect) or because patients of low SES are more often admitted to hospitals with higher overall complication rates (hospital-level effect). **Identifying the relative contribution of the hospital level is important in considering whether a factor should be included in risk adjustment; if an effect is primarily a hospital-level effect, adjusting for it is equivalent to adjusting for differences in hospital quality.** Thus, as an additional step, we assessed whether there was a "contextual effect" at the hospital level. To do this, we performed a decomposition analysis to assess the independent effects of the SES and race variables at the patient level and the hospital level. If, for example, all the elevated risk of complications for patients of low SES were due to lower quality/higher complication risk in hospitals with more patients of low SES, then a significant hospital-level effect would be expected with little-to-no patient-level effect. However, if the increased complication risk were solely related to higher risk for patients of low SES regardless of hospital effect, then a significant patient-level effect would be expected and a significant hospital-level effect would not be expected.

Specifically, we modelled each of the SES and race variables as follows: Let X_{ij} be a binary indicator of the SES or race status of the i^{th} patient at the j^{th} hospital, and X_j the percent of patients at hospital j with $X_{ij} = 1$. Then we added both $X_{ij} \equiv X_{patient}$ and $X_j \equiv X_{hospital}$ to the model. The first variable, $X_{patient}$, represents the effect of the risk factor at the patient level (sometimes called the "within" hospital effect), and the second variable, $X_{hospital}$, represents the effect at the hospital level (sometimes called the "between" hospital effect). By including both of these in the same model, we can assess whether these are independent effects, whether one effect dominates the other, or whether only one of these effects contributes. This analysis allows us to simultaneously estimate the independent effects of: 1) hospitals with higher or lower proportions of low SES patients or African-American patients on the complication rate of an average patient; and 2) a patient's SES or race on their own complication rates when seen at an average hospital. It is very important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect.

It is also important to note that the patient-level and hospital-level coefficients cannot be quantitatively compared because the patient's SES circumstance or race in the model is binary whereas the hospitals' proportion of low SES patients or African-American patients is continuous. Therefore, in order to quantitatively compare the relative size of the patient and hospital effects, we calculated a range of predicted probabilities of complication based on the fitted model.

Specifically, to estimate an average hospital effect, we calculated the predicted probabilities for the following scenarios: (1) Assuming all patients do not have the risk factor (X_{ij} =0) and hospital level risk factor is at 5% percentile (P5) of all hospital values; (2) Assuming all patients do not have the risk factor and hospital level risk factor is at 95% percentile (P95); (3) Assuming all patients do have the risk factor (X_{ij} =1) and hospital level risk factor is at 5% percentile (P5); (4) Assuming all patients have the risk factor and hospital level risk factor is at 95% percentile (P95). The average hospital effect is estimated by ((2)-(1) + (4)-(3))/2 (P95-P5). Then, to estimate an average patient effect, we first calculated the predicted probabilities by assuming patient-level risk factor equal to 0 or 1 at different hospital risk factor percentiles (0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%). Then at each of those percentiles, we could obtain the difference of predicted probabilities between all patients not having the risk factor and then all patients having the risk factor. We calculated the average of those differences in predicted probabilities ('delta') as the patient effect.

In summary, the difference in predicted probabilities at the 95th and 5th percentiles (P95-P5) estimates the hospital-level effect of the SES or race risk factor on complication. The difference in predicted probabilities when all patients have and do not have the SES or race risk factor (delta) estimates the patient-level effect of the SES or race risk factor on complication. The hospital-level effect is greater than the patient-level effect when P95-P5 is greater than delta. We used P95 and P5 rather than the maximum (P100) and minimum (P0) to avoid outlier values.

We also performed the same analysis for several clinical covariates to contrast the relative contributions of patient- and hospital-level effects of clinical variables to the relative contributions for the SES and race variables.

References:

Alter DA, Franklin B, Ko DT, et al. Socioeconomic status, functional recovery, and long-term mortality among patients surviving acute myocardial infarction. PloS one 2014; 8:e65130.

Blum, A. B., N. N. Egorova, E. A. Sosunov, A. C. Gelijns, E. DuPree, A. J. Moskowitz, A. D. Federman, D. D. Ascheim and S. Keyhani. "Impact of Socioeconomic Status Measures on Hospital Profiling in New York City." Circ Cardiovasc Qual Outcomes 7, no. 3 (2014): 391-7.

Bozic KJ, Grosso LM, Lin Z, et al. Variation in Hospital-Level Risk-Standardized Complication Rates Following Elective Primary Total Hip and Knee Arthroplasty. The Journal of Bone & Joint Surgery. 2014;96:640-647

Browne JA, Novicoff WM, D'Apuzzo MR. Medicaid payer status is associated with in-hospital morbidity and resource utilization following primary total joint arthroplasty. The Journal of bone and joint surgery. American volume. 2014;96(21):e180.

Cram P, Vaughan-Sarrazin MS, Wolf B, Katz JN, Rosenthal GE. A comparison of total hip and knee replacement in specialty and general hospitals. The Journal of bone and joint surgery. American volume. 2007;89(8):1675-1684.

Ibrahim SA. Racial variations in the utilization of knee and hip joint replacement: an introduction and review of the most recent literature. Current orthopaedic practice 2010;21:126-131

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and Medicaid patients. Health Affairs 2011; 30:1904-11.

Katz JN, Bierbaum BE, Losina E. Case mix and outcomes of total knee replacement in orthopaedic specialty hospitals. Medical care. 2008;46(5):476-480.

Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. The Journal of bone and joint surgery American volume 2003;85-a:27-32

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

Ong KL, Kurtz SM, Lau E, Bozic KJ, Berry DJ, Parvizi J. Prosthetic joint infection risk after total hip arthroplasty in the Medicare population. The Journal of arthroplasty. 2009;24(6 Suppl):105-109.

Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.

Skinner J, Chandra A, Staiger D, Lee J, McClellan M. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. Circulation 2005; 112:2634-41.

Taksler GB, Keating NL, Cutler DM. Explaining racial differences in prostate cancer mortality. Cancer 2012; 118:4280-9.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Below is a table showing the final variables in the model with associated odds ratios (OR).

Final Model Variables (variables meeting criteria in field 2b4.3) (**Dataset 1**)

Variable	04/2011-03/2014 OR (95% CI)
Age minus 65 (years above 65, continuous)	1.0 (1.02-1.03)
Male	1.2 (1.12-1.18)

Variable	04/2011-03/2014 OR (95% CI)
Index admissions with an elective THA procedure	1.4 (1.35-1.42)
Number of procedures (two vs. one)	1.7 (1.58-1.82)
Morbid obesity (ICD-9 code 278.01)	1.6 (1.52-1.66)
Other congenital deformity of hip (joint) (ICD-9 code 755.63)	1.1 (0.83-1.39)
Post traumatic osteoarthritis (ICD-9 codes 716.15, 716.16)	1.0 (0.82-1.18)
Metastatic cancer or acute leukemia (CC 7)	1.1 (0.94-1.26)
Cancer (CC 8-10)	0.9 (0.91-0.98)
Respiratory/heart/digestive/urinary/other neoplasms (CC 11-13)	0.9 (0.91-0.97)
Diabetes mellitus (DM) or DM complications (CC 15-20, 119-120)	1.2 (1.12-1.18)
Protein-calorie malnutrition (CC 21)	2.8 (2.59-3.02)
Bone/joint/muscle infections/necrosis (CC 37)	1.1 (1.07-1.21)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)	1.1 (1.09-1.18)
Osteoarthritis of hip or knee (CC 40)	1.0 (0.91-1.04)
Osteoporosis and other bone/cartilage disorders (CC 41)	1.0 (0.97-1.03)
Dementia or other specified brain disorders (CC 49-50)	1.2 (1.13-1.25)
Major psychiatric disorders (CC 54-56)	1.4 (1.29-1.43)
Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)	1.2 (1.10-1.29)
Cardio-respiratory failure or shock (CC 79)	1.2 (1.08-1.23)
Coronary atherosclerosis or angina (CC 83-84)	1.3 (1.30-1.37)
Stroke (CC 95-96)	1.1 (1.05-1.21)
Vascular or circulatory disease (CC 104-106)	1.1 (1.11-1.17)
Obstructive pulmonary disease (COPD) (CC 108)	1.5 (1.46-1.55)
Pneumonia (CC 111-113)	1.2 (1.19-1.31)
Pleural effusion/pneumothorax (CC 114)	0.9 (0.83-0.98)
Dialysis status (CC 130)	1.5 (1.25-1.79)
Renal failure (CC 131)	1.2 (1.15-1.24)
Decubitus ulcer or chronic skin ulcer (CC 148-149)	1.3 (1.21-1.36)
Trauma (CC 154-156, 158-161)	1.2 (1.14-1.25)
Vertebral fractures (CC 157)	1.1 (1.00-1.21)
Other injuries (CC 162)	1.1 (1.05-1.10)
Major complications of medical care and trauma (CC 164)	1.2 (1.16-1.29)

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Variation in Prevalence of the Factor across Measured Entities

The prevalence of SES factors and African-American patients in the total hip arthroplasty/total knee arthroplasty (THA/TKA) cohort varies across measured entities. The median percentage of dual eligible patients is 6.7% (interquartile range [IQR] 3.9%- 11.7%). The median percentage of African-American patients is 2.1% (IQR 0.0%- 6.3%). The median percentage of patients with an AHRQ SES Index score adjusted for cost of living at the census block group level equal to or below 42.7 is 12.9% (IQR 6.4%- 24.0%).

Empirical Association with the Outcome (Univariate)

The patient-level observed (unadjusted) THA/TKA complication rate is higher for dual eligible patients, 4.3%, compared with 3.1% for all other patients. The complication rate for African-American patients was also higher at 3.5% compared with 3.1% for patients of all other races. Similarly the complication rate for patients with an AHRQ SES index score equal to or below 42.7 was 3.5% compared with 3.1% for patients with an AHRQ SES index score above 42.7.

Incremental Effect of SES Variables and Race in a Multivariable Model

We then examined the strength and significance of the SES variables and race in the context of a multivariable model. Consistent with the above findings, when we include any of these variables in a multivariable model that includes all of the claims-based clinical variables, the effect size of dual eligibility is moderate (dual eligibility OR 1.21 [95% CI 1.16-1.27]) and the effect sizes of low AHRQ SES Index and African-American race are small (race OR 1.07 [95% CI 1.01-1.13], AHRQ SES Index OR 1.07 [95% CI 1.03-1.11]). The c-statistic is unchanged with the addition of any of these variables into the model. Furthermore the addition of any of these variables into the model has little to no effect on hospital performance. We examined the change in hospitals' RSCRs with the addition of any of these variables. The median absolute change in hospitals' RSCRs when adding a dual eligibility indicator is 0.0253% (IQR 0.0091% – 0.0444%, minimum -0.1884% – maximum 0.1242%) with a correlation coefficient between RSCRs for each hospital with and without dual eligibility added of 0.9987. The median absolute change in hospitals' RSCRs when adding a race indicator is 0.0252% (IQR 0.0136% – 0.0374%, minimum -0.0773% – maximum 0.1124%) with a correlation coefficient between RCRRs for each hospital with and without race added of 0.9996. The median absolute change in hospitals' RSCRs when adding an indicator for a low AHRQ SES index score adjusted for cost of living at the census block group level is 0.0237% (IQR 0.0114% – 0.0396%, minimum -0.0335% – maximum 0.1226%) with a correlation coefficient between RSCRs for each hospital with and without an indicator for a low AHRQ SES index score added of 0.9993.

Contextual Effect Analysis

As described in 2b4.3, we performed a decomposition analysis for each SES and race variable to assess whether there was a corresponding contextual effect. In order to better interpret the magnitude of results, we performed the same analysis for selected clinical risk factors. The results are described in the first table below (the decomposition table).

Both the patient-level and hospital-level dual eligible and low AHRQ SES Index effects were significantly associated with hip/knee complications in the decomposition analysis. The patient-level race effect was not appreciably different from zero, though the hospital-level effect was significant. **That the hospital level effects were significant indicates that if the dual eligible or low AHRQ SES Index variables are used in the model to adjust for patient-level differences, then some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality**.

To assess the relative contributions of the patient- and hospital-level effects, we calculated a range of predicted probabilities of complication for the SES or race variables and clinical covariates (comorbidities), as described in section 2b4.3. The results are presented in the figure and second table below (table of predicted probabilities for SES and race variables).

For low AHRQ SES and race variables, the hospital-level effect (P95-P5) is greater than the patient-level effect (delta). The hospital-level effect is approximately equal to the patient-level effect for dual eligibility (second table below; the table of predicted probabilities for SES and race variables). For clinical variables, the patient-level effect (delta) is greater than the hospital-level effect (P95-P5) for and COPD and for metastatic cancer, although the patient-level effect for metastatic cancer was not statistically significant likely due to sample size (third table below; the table of predicted probabilities for clinical variables). The hospital-level and patient-level effects for renal failure appear equivalent. This pattern demonstrates that SES and race variables have a much greater hospital-level effect than patient-level effect. The clinical variables had the opposite pattern, with a greater effect at the patient level than at the hospital level. Therefore, including SES and race variables into the model would predominantly adjust for a hospital-level effect, which is an important signal of hospital quality.

In the context of our conceptual model, we find clear evidence supporting the first two mechanisms by which SES might be related to poor outcomes. First we find that, although unadjusted rates of complication are higher for patients of low SES and African-American race, the addition of SES to the complication risk model, which already adjusts for clinical factors, makes very little difference. In particular, the Odds Ratios associated with each variable in the multivariate models are small and there is little-to-no change in model performance or hospital results with the addition of SES. This suggests that the model already largely accounts for the differences in clinical risk factors (degree of illness and comorbidities) among patients of varied SES.

Second, the predominance of the hospital-level effect of SES and race variables in the decomposition analyses suggests the risk associated with low SES is in large part due to lower quality of care at hospitals where more patients with these risk factors are treated; hospitals caring for socially- and economically-disadvantaged patients have higher complication risk for **all** of their patients. Patients with low SES indicators or African-American patients tend to receive care more frequently at lower quality hospitals compared with patients with high SES indicators. Direct adjustment for patient SES would essentially “over adjust” the measure, that is to say, it would be adjusting for an endogenous factor, one that influences the outcome through the site of treatment (hospital), as much as through an attribute of the patient.

In comparison, we did not observe the same predominance of the hospital-level effect among the clinical covariates, reinforcing the sense that SES and race factors have a distinct causal pathway in their impact on complication risk.

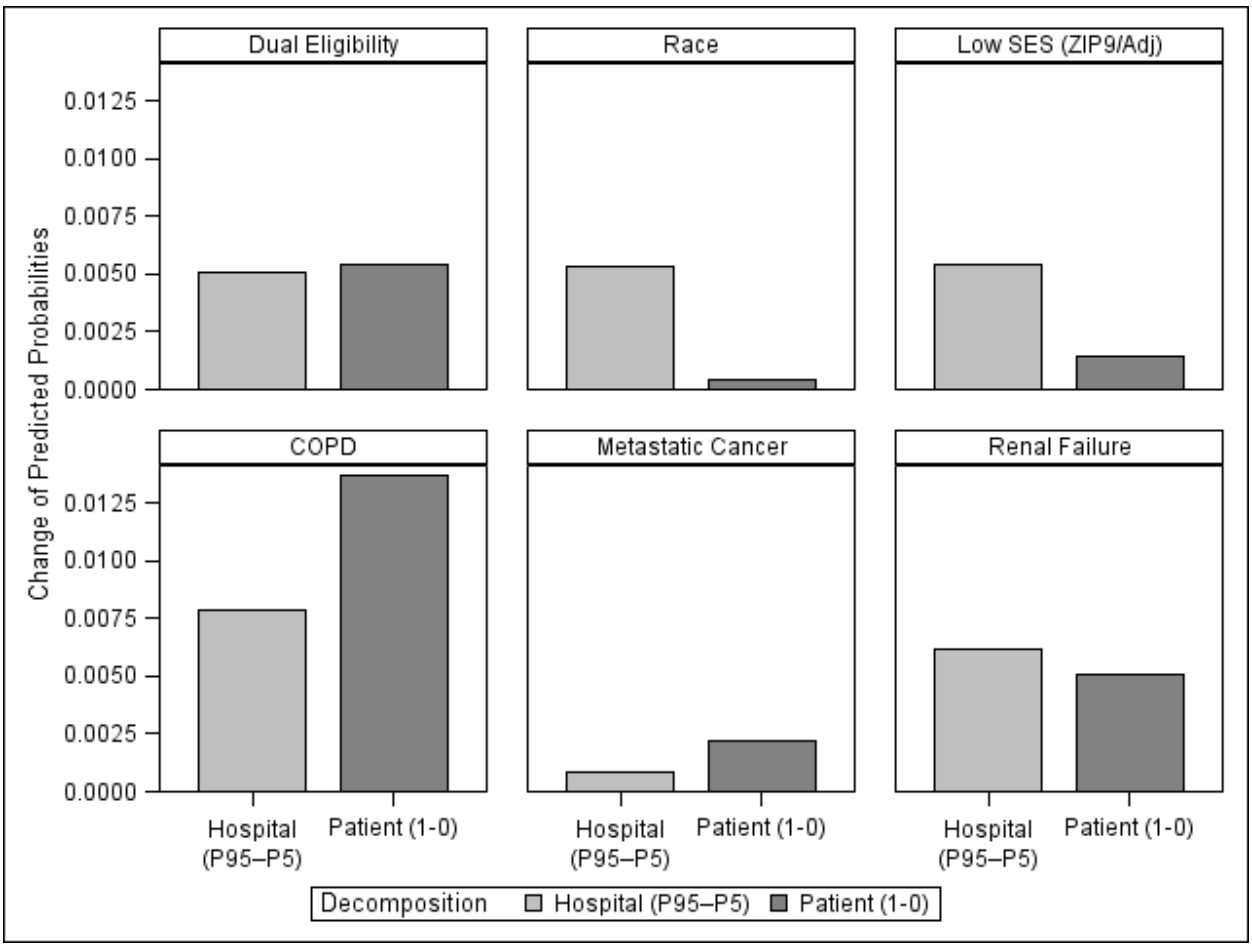
Summary

We found wide variation in the distribution of the three SES and race factors we examined, and we found that all three had some association with complication risk. However, adjustment for these factors did not have an appreciable impact on hospital RSCRs, suggesting that existing clinical risk factors capture much of the risk related to low SES and African-American race. More importantly, we found that for all three factors there was an equal (for dual eligible status) or greater (for AHRQ SES Index score) hospital-level effect, compared with the patient-level effect, indicating that patient-level adjustment alone would adjust for quality differences between hospitals. Therefore, we did not include SES factors in our final risk model.

THA/TKA Complication Decomposition Analysis

Parameter	Estimate (Standard Error)	P-value
Dual Eligible – Patient-Level	0.1627 (0.0226)	<0.0001
Dual Eligible – Hospital-Level	0.5068 (0.0903)	<0.0001
African-American – Patient-Level	0.0141 (0.0295)	0.6315
African-American – Hospital-Level	0.7766 (0.1071)	<0.0001
Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)– Patient-Level	0.0451 (0.0182)	0.0133
Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)– Hospital-Level	0.3584 (0.0690)	<0.0001
Renal Failure – Patient-Level	0.1698 (0.0192)	<0.0001
Renal Failure – Hospital-Level	1.6004 (0.2414)	<0.0001
Metastatic Cancer – Patient-Level	0.0763 (0.0751)	0.3094
Metastatic Cancer – Hospital-Level	1.6315 (0.7821)	0.0370
COPD – Patient-Level	0.3956 (0.0157)	<0.0001
COPD – Hospital-Level	1.0763 (0.1432)	<0.0001

Change of Predicted Probabilities for SES and Race Compared with Clinical Variables in THA/TKA Complication Measure



*Low SES (ZIP9/Adj) measured by linking patients' 9-digit ZIP codes to AHRQ SES Index at the census block group level, adjusted for cost of living

Predicted Probabilities for SES and Race Variables in the THA/TKA Complication Measure

Hospital SES/Race Risk Factor Percentile	Dual Eligibility				Race				Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)			
	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)
0%	0.0000	0.0296	0.0345	0.0050	0.0000	0.0299	0.0303	0.0004	0.0000	0.0294	0.0307	0.0013
5%	0.0159	0.0298	0.0348	0.0050	0.0000	0.0299	0.0303	0.0004	0.0132	0.0295	0.0308	0.0013
10%	0.0230	0.0299	0.0349	0.0050	0.0000	0.0299	0.0303	0.0004	0.0272	0.0297	0.0310	0.0013
25%	0.0392	0.0301	0.0352	0.0051	0.0000	0.0299	0.0303	0.0004	0.0638	0.0300	0.0314	0.0013
50%	0.0667	0.0305	0.0356	0.0051	0.0206	0.0303	0.0308	0.0004	0.1293	0.0307	0.0321	0.0013
75%	0.1170	0.0313	0.0365	0.0052	0.0631	0.0313	0.0317	0.0004	0.2400	0.0319	0.0333	0.0014
90%	0.2186	0.0329	0.0383	0.0055	0.1422	0.0332	0.0337	0.0004	0.3898	0.0336	0.0350	0.0015
95%	0.3208	0.0345	0.0403	0.0057	0.2204	0.0352	0.0357	0.0005	0.4966	0.0348	0.0363	0.0015
100%	0.9294	0.0462	0.0537	0.0076	0.9380	0.0593	0.0600	0.0008	0.9512	0.0406	0.0424	0.0018
P95 – P5 (Hospital Effect)	-	0.0047	0.0055	-	-	0.0053	0.0054	-	-	0.0053	0.0055	-

Predicted Probabilities for Clinical Variables in the THA/TKA Complication Measure

Hospital SES/Race Risk Factor Percentile	Renal Failure				Metastatic Cancer				Chronic Obstructive Pulmonary Disease			
	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)
0%	0.0000	0.0246	0.0290	0.0044	0.0000	0.0285	0.0307	0.0022	0.0000	0.0250	0.0365	0.0115
5%	0.0368	0.0261	0.0307	0.0046	0.0000	0.0285	0.0307	0.0022	0.0674	0.0268	0.0391	0.0123
10%	0.0484	0.0265	0.0312	0.0047	0.0000	0.0285	0.0307	0.0022	0.0825	0.0272	0.0397	0.0125
25%	0.0665	0.0273	0.0321	0.0048	0.0000	0.0285	0.0307	0.0022	0.1083	0.0279	0.0407	0.0128
50%	0.0862	0.0281	0.0331	0.0050	0.0035	0.0287	0.0308	0.0022	0.1410	0.0289	0.0421	0.0132
75%	0.1128	0.0293	0.0345	0.0052	0.0078	0.0289	0.0311	0.0022	0.1855	0.0302	0.0441	0.0138
90%	0.1416	0.0306	0.0360	0.0054	0.0133	0.0291	0.0313	0.0022	0.2353	0.0318	0.0463	0.0145
95%	0.1644	0.0317	0.0373	0.0056	0.0185	0.0294	0.0316	0.0022	0.2781	0.0333	0.0484	0.0151
100%	0.4000	0.0454	0.0532	0.0078	0.0741	0.0320	0.0344	0.0024	0.9333	0.0645	0.0921	0.0277
P95 – P5 (Hospital Effect)	-	0.0057	0.0066	-	-	0.0008	0.0009	-	-	0.0065	0.0093	-

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Approach to assessing Model Performance (**Dataset 1, Dataset 2, and Dataset 3**)

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the cohorts:

Discrimination Statistics

(1) Area under the receiver operating characteristic (ROC) curve (or the c-statistic) (the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects. Therefore, we would hope to see a wide range between the lowest decile and highest decile)

Calibration Statistics

(3) Over-fitting indices (the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients)

We tested the performance of the model for **Dataset 1** (current public reporting cohort) and **Dataset 3** (development dataset) described in section 1.7. During initial measure development, we tested the performance of the model developed in a randomly selected half of the hospitalizations for THA/TKA in 2008 compared with performance calculated from hospitalizations from the other half (**Dataset 3**). As a part of measure reevaluation, each year we assess temporal trends in model performance in the combined 3-year public reporting data (**Dataset 1**). Below, we report the model performance only for the 3-year combined results. For results for each individual year within the combined 3-year data please see the attached specifications report.

Reference:

F.E. Harrell and Y.C.T. Shih, Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

For the development cohort (**Dataset 3**) the results are summarized below:

First half of randomly split development sample: C-statistic = 0.69; Predictive ability (lowest decile %, highest decile %) = (2, 15)
Second half of randomly split development sample: C-statistic = 0.70; Predictive ability (lowest decile %, highest decile %) = (2, 15)

For the current measure cohort (**Dataset 1**) the results are summarized below:

C statistic =0.65; Predictive ability (lowest decile %, highest decile %) = (1.5, 7.4)

For comparison of model with and without inclusion of SDS factors, see Section 2b4.4b.

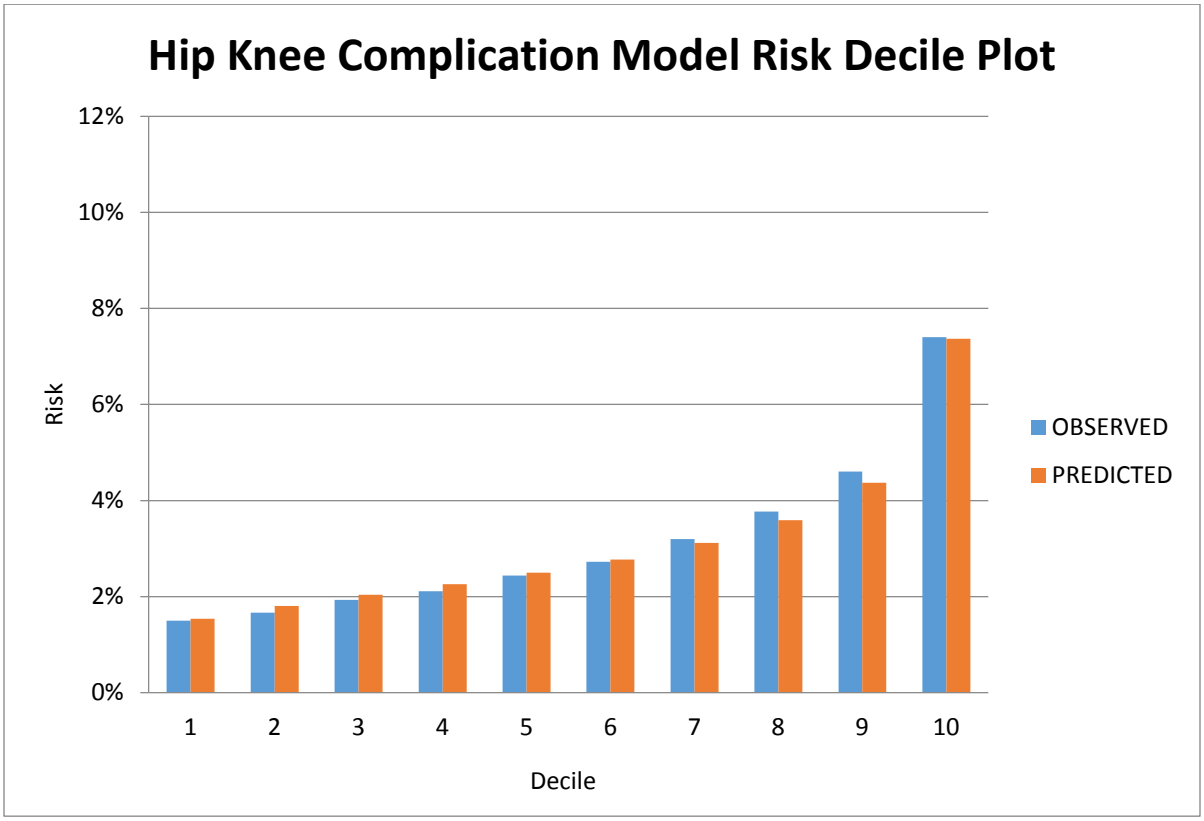
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

For the original measure development cohort (**Dataset 3**) the results are summarized below:

First half of split sample: Calibration: (0, 1)
Second half of split sample: Calibration: (0.04, 1.02)

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from April 2011 to March 2014 (**Dataset 1**).



2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted*)

Discrimination Statistics

The C-statistics of 0.69, 0.70, and 0.65 for the model development, validation, and current public reporting data (**Datasets 3** first half of split sample**, 3** second half of split sample**, and 1** respectively) demonstrate consistent and good model discrimination (**Datasets 1 and 3**). The models also indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

Calibration Statistics

Over-fitting (Calibration γ0, γ1)

If the γ0 in the development and validation samples (**Dataset 3**) are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration value of close to 0 at one end and close to 1 at the other end indicates good calibration of the model.

Risk Decile Plots

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates excellent discrimination of the model and good predictive ability.

Overall Interpretation

Interpreted together, our diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Testing of Measure in All-Payer Cohort (Dataset 4)

We tested the THA/TKA complication measure in an all-payer patient population of adults aged 18 years and older so that it can be applied to both Medicare and all-payer populations. Using data from California as well as CMS Medicare FFS data for California hospitals, we performed analyses to determine whether the THA/TKA complication measure can be applied to all adult patients, including FFS Medicare patients aged 65+, non-FFS Medicare patients aged 65+, and patients aged 18-64 years at the time of admission. The THA/TKA complication model developed in Medicare FFS 65+ patients uses inpatient and outpatient data for risk adjustment (consistent with CMS' publicly reported mortality and readmission measures for AMI, heart failure, and pneumonia).

We applied the model to all-payer data from California. The analytic sample included 59,828 cases aged 18 and older in the 2006 California Patient Discharge Data. When used in all-payer data, only admission claims data are used for risk adjustment, as the hospital discharge databases do not have outpatient claims.

To address the question of how well the models perform when applied to all patients 18+, we used the California Patient Discharge Data (PDD). Specifically, using 2006 data, we created measure cohorts with up to one year of hospital inpatient claims history and 30-day follow-up data. For the THA/TKA complication measure, we:

1. Created the patient cohorts using the respective measure inclusion and exclusion criteria (with the exception of including all patients 18+), and compared the FFS 65+, non-FFS 65+, and 18-64 year-old patient subgroups with respect to the distributions of risk factors and the crude outcome rates.

2. Fit the models in all patients 18+ and: (i) examined overall model performance in terms of the C statistic, (ii) compared performance (C statistic and predictive ability) across the patient subgroups (FFS 65+, non-FFS 65+, all 65+, and all-payer 18-64), and (iii) compared the distribution of Pearson residuals (model fit) across the patient subgroups.

3. Fit the models separately in each patient subgroup and compared ORs associated with the risk factors to assess differences in magnitude or direction of ORs among the subgroups.

To determine whether the relationship between each risk factor and the outcome differed for those aged 65+ vs. 18-64 in ways that would affect measure results, we:

1. Fit the models in all patients 18+ and tested interaction terms between age (65+ vs. 18-64) and each of the other risk factors.

2. Fit the models in all patients 18+ with interaction terms and compared performance (c-statistic and predictive ability) across the patient subgroups.

3. Fit the models in all patients 18+ with and without interaction terms and (i) conducted a reclassification analysis to compare risk prediction at the patient level; (ii) compared the c-statistic; and (iii) compared hospital-level risk-standardized rates using a scatterplot and the ICC to assess whether the models with interactions are statistically different from the current models in profiling hospital rates.

All patient-level models were estimated using a logistic regression model.

Results of Measure Testing in All-Payer Cohort

When the model was applied to all patients 18 and over (18+), overall discrimination was good (c-statistic=0.69). In addition, there was good discrimination and predictive ability in both those aged 18-64 and those aged 65+. Moreover, the distribution of Pearson residuals was comparable across the patient subgroups. When comparing the model with and without interaction terms, (a) the reclassification analysis demonstrated good patient-level risk prediction (12.0% to 44.1% vs. 13.0% to 43.2%, respectively, from the bottom decile to the top decile of the prediction values); (b) the c-statistic was nearly identical (0.69 vs. 0.69); and (c) hospital-level risk-standardized rates were highly correlated (r=0.998; ICC=0.996). Thus, the inclusion of the interactions did not substantively affect either patient-level model performance or hospital-level results

There are some differences in the risk factor profiles and crude outcome rate among patient subgroups. In general, the prevalence of risk factors was similar in FFS 65+ and non-FFS 65+ patients. There were slight differences in the prevalence of two risk factors in the complication measure (Osteoporosis and Other Bone/Cartilage Disorders [CC 41] and Chronic Atherosclerosis [CC 83-84]) When comparing risk factor prevalence estimates between those 65+ and younger patients aged 18-64, frequencies were generally either lower in the younger cohort or similar between the groups. For some risk factors, including having an index elective THA procedure and morbid obesity (ICD-9 code 278.01) in the complication model, prevalence estimates were in fact higher in younger than in older patients. The complication rates were very similar between older patients 65+ and younger patients 18-64 years. Odds Ratios were generally similar for FFS 65+ and non-FFS 65+ patients, although for some risk factors, there were differences in magnitude of effect between younger and older patients.

For the complication model, there were significant age-by-risk-factor interaction terms for two variables (Older and Male, and Older and Major Complications of Medical Care and Trauma [CC 164]). Inclusion of the interaction terms, however, did not substantively change the level of discrimination and predictive ability across the patient subgroups.

In addition, when comparing patient risk classifications for the measure with and without interaction terms, the reclassification analysis demonstrated good patient-level risk prediction: for all patient subgroups, nearly 100% of patients were in a similar risk category (defined as being in the same or adjacent category) regardless of risk-adjustment strategy. Moreover, the C-statistic was identical (0.68) for the models with and without interaction terms for THA/TKA complication. Finally, when comparing the measure with and without interaction terms, the hospital-level risk-standardized rates estimated by the two versions of the model were highly correlated (r = 0.998 for THA/TKA complication).

This measure is fully risk-adjusted using a hierarchical logistic regression model to calculate hospital risk-standardized complication rates (RSCRs) accounting for differences in hospital case mix.

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

For public reporting of the measure, CMS characterizes the uncertainty associated with the RSCR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSCR's interval estimate does not include the national observed complication rate (is lower or higher than the rate), then CMS is confident that the hospital's RSCR is different from the national rate, and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate," respectively. If the interval includes the national rate, then CMS describes the hospital's RSCR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 THA/TKA procedures in the three-year period.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Analyses of Medicare FFS data show variation in RSCRs among hospitals. Using data from April 2011 – March 2014 (**Dataset 1**), the median hospital RSCR was 3.1%, with a range of 1.4% to 6.9%. The interquartile range was 2.9%-3.4%.

Out of 3,507 hospitals in the U.S., 54 performed "better than the U.S. national rate," 2,711 performed "no different from the U.S. national rate," and 45 performed "worse than the U.S. national rate." 697 were classified as "number of cases too small" (fewer than 25) to reliably tell how well the hospital is performing.

Note that this analysis included index admissions from April 2011 – March 2014 from the 2015 public reported data (**Dataset 1**).

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

There has been a downward trend in complication rates between 2011 and 2014. However, the rate following THA/TKA remains higher than expected for an elective procedure at 2.95% in 2013-2014. The variation in rates suggests there are meaningful differences across hospitals in complications following THA/TKA admissions.

Note: From the April 2011 to March 2012 reporting year to the April 2013 to March 2014 reporting year, the observed THA/TKA complication rate decreased from 3.34% (April 2011 – March 2012) to 2.95% (April 2013 – March 2014). The observed complication rate for the 3-year combined public reporting period (April 2011 – March 2014) for THA/TKA Medicare FFS patients is 3.14%.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

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

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and*

compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). ***Comparability is not required when comparing performance scores with and without SDS 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.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

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

Administrative data are routinely collected as part of the billing process.

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

There are no fees associated with the use of this measure.

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Planned	Current Use (for current use provide URL)
Quality Improvement with Benchmarking (external benchmarking to multiple organizations)	Public Reporting Hospital Inpatient Quality Reporting (IQR) Program http://cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html

4a.1. For each CURRENT use, checked above, provide:

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

Public Reporting

Program Name, Sponsor: Hospital Inpatient Quality Reporting (Hospital IQR) Program, Centers for Medicare and Medicaid Services (CMS)

Purpose: The Hospital IQR program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates. Initially, the MMA provided for a 0.4 percentage point reduction in the annual market basket (the measure of inflation in costs of goods and services used by hospitals in treating Medicare patients) update for hospitals that did not successfully report. The Deficit Reduction Act of 2005 increased that reduction to 2.0 percentage points.

In addition to giving hospitals a financial incentive to report the quality of their services, the Hospital IQR program provides CMS with data to help consumers make more informed decisions about their health care. Some of the hospital quality of care information gathered through the program is available to consumers on the Hospital Compare website at: www.hospitalcompare.hhs.gov.

Geographic area and number and percentage of accountable entities and patients included:
The Hospital IQR program includes all Inpatient Prospective Payment System (IPPS) non-federal acute care hospitals and VA hospitals in the United States. The number and percentage of accountable hospitals included in the program, as well as the number of patients included in the measure, varies by reporting year. For 2015 public reporting, the RSCR was reported for 3,507 hospitals across the U.S. The final index cohort includes 892,455 admissions.

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

N/A. This measure is currently publicly reported.

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

N/A. This measure is currently publicly reported.

4b. Improvement

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

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

There has been progress in reducing the RSCR following THA/TKA. The median RCRR decreased by 0.4 absolute percentage points from April 2011-March 2012 (median RSCR: 3.3%) to April 2013-March 2014 (median RSCR: 2.9%). The median hospital RSCR from April 2011-March 2014 was 3.1% (IQR 2.9% - 3.4%).

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We did not identify any unintended consequences during measure development, model testing, or re-specification. However, we are committed to monitoring this measure's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care, increased patient morbidity and mortality, and other negative unintended consequences for patients.

5. Comparison to Related or Competing Measures

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

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

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB).
0564 : Cataracts: Complications within 30 Days Following Cataract Surgery Requiring Additional Surgical Procedures
1551 : Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
2052 : Reduction of Complications through the use of Cystoscopy during Surgery for Stress Urinary Incontinence

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

5a. Harmonization

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

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

Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
We did not include in our list of related measures any non-outcome measures (for example, process measures) with the same target population as our measure. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

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

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Lein, Han, Lein.han@cms.hhs.gov, 410-786-0205-
Co.3 **Measure Developer if different from Measure Steward:** Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/ CORE)
Co.4 **Point of Contact:** Karen, Dorsey, karen.dorsey@yale.edu, 203-764-5700-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The working group involved in the initial measure development is detailed in the original technical report available at www.qualitynet.org.

Our measure development team consisted of the following members:
Laura M. Grosso, PhD, MPH
Jeptha P. Curtis, MD
Zhenqiu Lin, PhD
Lori L. Geary, MPH
Smitha Vellanky, MSc
Carol Oladele, MPH
Yongfei Wang, MS
Elizabeth E. Drye, MD, SM
Harlan M. Krumholz, MD, SM

Working Group Members:
Daniel J. Berry, MD
Kevin J. Bozic, MD, MBA
Robert Bucholz, MD
Lisa Gale Suter, MD
Charles M. Turkelson, PhD
Lawrence Weis, MD

Technical Expert Panel Members:
Mark L. Francis, MD
Cynthia Jacelon, PhD, RN, CRRN
Norman Johanson, MD
C. Kent Kwoh, MD
Courtland G. Lewis, MD
Jay Lieberman, MD
Peter Lindenauer, MD, M.Sc.
Russell Robbins, MD, MBA
Barbara Schaffer
Nelson SooHoo, MD, MPH
Steven H. Stern, MD
Richard E. White, Jr., MD

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2013
Ad.3 **Month and Year of most recent revision:** 06, 2015
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** N/A
Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 2998
De.2. **Measure Title:** : Infection rate of bicondylar tibia plateau fractures
Co.1.1. **Measure Steward:** Orthopedic Trauma Association
De.3. **Brief Description of Measure:** Percent of patients aged 18 years and older undergoing ORIF of a bicondylar tibial plateau fracture who develop a postoperative deep incisional wound infection based on CDC guidelines for deep infection associated with implants
1b.1. **Developer Rationale:** Bicondylar tibial plateau are difficult injuries to treat and are often complicated by infection, nonunion, and compartment syndrome. Infection rate of these injuries is reported to be between 20-30% at high volume centers with experienced surgeons following a staged fixation protocol with dual incisions. The lowest infection rate ever reported for bicondylar tibia plateau fractures treated with ORIF is 8%. As orthopedic surgeons the OTA hopes that we can "do better than this" Routine reporting of infection rates associated with these injuries can help all surgeons evaluate their results and lead to improved Quality Improvement processes to drive down the current infection rates. These injuries were chosen because they are commonly treated by orthopedic surgeons with advanced Trauma training and they have some of the highest reported rates of this infection of any operation. Infection increases the cost of care as treatment involves multiple debridements, multiple clinic and hospital visits and often increases the length of stay during inpatient care. This long treatment course also immobilizes the patient which affects their physical and emotional health. Orthopedic surgeons cannot change the injury or patient, but can attempt to optimize modifiable patient characteristics, control timing and type of fixation, surgical technique and post op protocols for treatment of these injuries. No one action will likely be detectable, but a series of actions and processes that are in control of the surgeon can be implemented that may decrease the high infection rates that are reported. Surgeons with high rates can re-evaluate their actions and processes and surgeons with low rates can publish their processes to encourage best practices by our colleagues.
A deep wound infection increases the morbidity of the injury and treatment on the patient. Infection raises the risk of nonunion and certainly increases number of surgeries required, pain suffered and intensity of interventions.

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.
Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.
Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.
Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.
Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

S.4. **Numerator Statement:** Number of patients aged 18 years and older undergoing ORIF of a bicondylar tibial plateau fracture who develop a postoperative deep incisional infection associated with an implant within 1 year of fracture fixation. We do not have adequate data to provide adequate risk stratification at this time.
S.7. **Denominator Statement:** All patients undergoing ORIF of a closed bicondylar tibial plateau fracture aged 18 years or older. Patients can be identified with either an ICD-10 code (S82.141, S82.142) or by CPT billing codes. (27536). Risk calculation can be added once adequate volume of patients are enrolled.
S.10. **Denominator Exclusions:** N/A

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Electronic Clinical Data: Registry, Other

S.26. **Level of Analysis:** Clinician: Group/Practice, Clinician: Individual, Facility

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

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

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

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** N/A

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

The developer stated that bicondylar tibial plateau fractures are difficult to treat and often complicated by infection that ranges from 20 – 30% at high volume centers, with experienced surgeons. Information is provided that the lowest infection rate reported for these fractures treated with open reduction and internal fixation (ORIF) is 8%. These surgeries have some of the highest reported infection rates of any operation; and they increase cost of care. A deep wound infection increases the morbidity of the injury and treatment on the patient. Infection raises the risk of nonunion and increases number of surgeries required, pain suffered, and intensity of interventions.

By reporting this infection rate, the developer opines that surgeons can/will take action to reduce infection rates and avoid infection sequelae or, in the case of surgeons with low infection rates, can disseminate information regarding processes that help keep rates low.

Question for the Committee:

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

Guidance from the Evidence Algorithm

Assessment of performance of health outcome (Box 1) → relationship between health outcome and action supported by rationale → Pass

Preliminary rating for evidence: X Pass ☐ No Pass

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

Maintenance measures – increased emphasis on gap and variation

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

The developer provided information that the infection rate for these fractures ranges from 20 – 30% and provided literature that reports a high rate of deep infection when treating bicondylar tibial plateau fractures. Barei, et al., identified an 8.4%(7/83) infection rate. Ruffolo, et al., retrospectively analyzing 140 bicondylar tibial plateau fractures, reported an infection rate of 23.6 % (33/140). Morris, et al., through a retrospective review of 302 fractures, reported a 14.2%(43/302) infection rate.

Disparities

The developer reports that disparities data are not yet available, noting that demographic and SES data can be collected through the Qualified Clinical Data Registry (QCDR) for later analysis.

Questions for the Committee:

- o *Is there a gap in care that warrants a national performance measure?*
- o *Are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: X High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

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

1a. Evidence to Support the Measure Focus

- Excellent measure. Very important concept and good gap.
  Seems easily defined in medical record, registry and EHR data
  Note that QCDR data WILL be publicly reported

- This is a health outcome measure. The rationale is collect data to allow surgeon comparison and potential improvement by reduction in infection rate. However, no clear interventions are available in the literature to reduce the rate.

1b. Performance Gap

- The performance gap is unknown. Small retrospective studies demonstrate a low of 8.4% (7/83 patients) to 23.6%. Rate is generally regarded as 20-30%.

  Performance gap, if any is unclear based on data presented.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Electronic clinical data/registry

Specifications:

- The measure assesses the number of patients (18 and older) undergoing ORIF for a closed bicondylar tibial plateau fracture who develop deep incisional infection associated with an implant within 1 year of fixation. Infection will be identified by an operative report for irrigation and debridement of the operative wound and confirmed culture-positive intraoperative findings. Patients can be identified with either an ICD-10 code (S82.141, S82.142) or by CPT billing codes. (27536) and have an admission for a post op wound infection (CPT 10180)
- Criteria for infection is drawn from CDC criteria for deep incisional infection with exception of the 1 year after index operative procedure criterion (CDC criterion is within 30 or 90 days).
- The denominator specifies that all patients (18 and older) who undergo the procedure for the specified fracture are included without exclusion. There is no risk adjustment or stratification.
- The level of analysis is clinician: group/practice and individual/facility and the care setting is hospital/acute care setting. Better quality = lower score.

Questions for the Committee:

- o *Are all the data elements clearly defined? Are all appropriate codes included?*
- o *Is the logic clear?*

- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

To demonstrate reliability of the measure, the developer presented information from a secondary evaluation of bicondylar tibial plateau fractures from two large studies for which it had access to patient data. Of the 440 patients in these studies, 77 were selected for further review based on the fact that the patients (23.6% of one study and 14.2% of the second study) were diagnosed with infected bicondylar tibial plateau fracture. Through radiographs and CT scans, all 77 were confirmed to be bicondylar tibial plateau fractures. Through review of operative reports for irrigation and debridement and organism positive laboratory data, 76 of the 77 fractures were confirmed to be infected for an agreement rate of 99.42%. The remaining patient from this group had a debridement of a fluid collection with negative culture.

Additionally, of those patients identified as having closed bicondylar tibia plateau fractures on x-ray with no evidence of deep infection, 95 were randomly selected and evaluated. All 95 patients were confirmed as having closed bicondylar tibial plateau fractures without infection based on lack of operative reports for irrigation and debridement and no laboratory data indicating presence of infection.

Agreement was found in 171 of 172 cases reviewed or 99.42% of observations with a Kappa of 0.988. Sensitivity = 100%; Specificity = 99%; Positive Predictive Value = 98.7%

Sensitivity measures the proportion of actual positives that are correctly identified as such.

Specificity measures the proportion of actual negatives that are correctly identified as such.

Positive predictive values are the proportions of positives and negatives that are true positive and true negatives.

This satisfies NQF requirements for data element reliability and validity testing.

Reliability testing level ☐ **Measure score** **X Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure **X Yes** ☐ **No**

Questions for the Committee:

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

Guidance from the Reliability Algorithm: Specifications precise (Box 1) – Testing of patient-level data (Box 3) – Testing with patient-level data (Box 10 Validity) – Method described and appropriate (Box 11) – Confidence that data are reliable (Box 12) Highest score = Moderate

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

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. **X Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Does the Committee agree that the specifications are consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** **X Data element testing against a gold standard** ☐ **Both**

Validity testing method and results: See Reliability section above.

Questions for the Committee:

- *Is the test sample and approach to testing adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

There are no exclusions.

Questions for the Committee:

- *Are there exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** **X None** ☐ **Statistical model** ☐ **Stratification**

The developer states that patient factors, injury factors and socioeconomic status have not been consistently associated with differences in surgical site infection (SSI) in patients with this surgery. Characteristics of the 43 patients with deep wound infection from one institution were further analyzed and a conclusion reached that there was no reason to believe that the demographics would be different in other institutions.

Questions for the Committee:

- *Is there any evidence that contradicts the developer's rationale and analysis for not risk adjusting the measure at this time?*
- *Is the Committee aware of risk factors that should be considered now?*
- *Is the plan for establishing risk factors for future incorporation in the measure reasonable?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

There are no data on which to consider question of meaningful difference. The developer plans to collect clinical and demographic data over a period of 3 years.

Question for the Committee:

- *Is the proposed plan reasonable?*

2b6. Comparability of data sources/methods:

There is one set of measure specifications. Data source is registry.

2b7. Missing Data

The developer stated there was no known missing data in the testing samples.

Guidance from the Validity Algorithm: Specifications precise (Box 1) – Potential threats (Box 2) – Testing with patient-level data (Box 10) – Method described and appropriate (Box 11) – Confidence that data are reliable (Box 12) Highest score = Moderate

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

Committee pre-evaluation comments

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

2a1. & 2b1. Specifications

- The number of patients undergoing ORIF for a closed bicondylar timbial plateau fracture who develop deep incision infection within 1 year. ICD-10 and CPT billing code for identification. Denominator is all ORIF for a closed bicondylar timbial plateau fracture. >/= 18 y/o

  high Kappa statistic

2a2. Reliability – Testing

- on retrospective review of 440 patients; 76 of 77 fractures were confirmed to be infected for agreement rate of 99.42%.

2b2. Validity – Testing

- data collection is voluntary;

Criterion 3. Feasibility

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

The developer states that data are generated or collected and used by healthcare personnel during the provision of care; some of the data elements are in defined field in electronic sources.

There are no costs for participating in the QCDR. Access is free for all Orthopedic Trauma Association (OTA) members. Data entry of the 20-30 data fields is the responsibility of participating sites.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- Data collection is voluntary. No incentive for reporting or consequence, but what is the benchmark?

Criterion 4: Usability and Use

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

Current uses of the measure

Publicly reported? ☐ **Yes X No**

Current use in an accountability program? ☐ **Yes X No**

OR

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

Accountability program details This is a new initiative for OTA, which has not had quality metrics for public reporting. The developer notes that the measure could also be used by the American College of Surgeons as a quality indicator for their trauma centers. The QCDR is not currently transparent to the public but could be made to be so. OTA could also publish overall infection rates and ranges so any surgeon or center could review their cases internally.

Improvement results N/A (new measure)

Unexpected findings (positive or negative) during implementation Testing involved review of medical charts of patients from 2 series of closed bicondylar tibia plateau fractures. There were no unexpected findings.

Potential harms None identified.

Feedback: None for this new orthopedic trauma measure.

***Questions for the Committee**:*

- o *Does the Committee envision approaches, not outlined by the developer, that should be considered for use of the performance results to further the goal of high-quality, efficient healthcare?*
- o *Do the benefits of the measure outweigh any potential unintended consequences?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- Maybe used to compare surgeon to surgeon. However no risk comparison is made between fracture cases. Usability is questionable.

Criterion 5: Related and Competing Measures

Related or competing measures NA

Harmonization NA

Pre-meeting public and member comments

- Smith & Nephew strongly supports quality measure #2998, titled "Infection rate of bicondylar tibia plateau fractures", as this measure would focus efforts around infection prevention and clinical protocols for this vulnerable patient group at high risk of infection. An infection rate reported to approach 30% is a significant burden. Efforts to lower this risk through mitigation of modifiable risk factors and application of evidence-based risk reduction strategies should be encouraged.

  One treatment strategy proven to mitigate infection risk in a level 1 study of tibial plateau fractures was negative pressure wound therapy (NPWT). In a prospective randomized trial of 263 fractures in 249 patients with tibial plateau, pilon and calcaneal fractures, patients randomized to NPWT experienced a

statistically significant reduction in infection rates (23 infections in control group vs. 14 in the treatment arm; P=.049) (Stannard et al, 2012). Of 117 tibial plateau fractures, the largest subgroup, there was a two-fold higher relative risk of infection in the control group; that is, infection was identified in 9/55 (16.3%) of control compared to 5/62 (8.1%) of NPWT treated fractures. Among all fractures, the relative risk of developing an infection was 1.9 times higher in the control group than in those treated with NPWT. Additionally, significantly fewer NPWT treated fractures experienced wound dehiscence after discharge compared to the control group, 20/122 (16.5%) compared to 12/141 (8.6%), respectively, and, there was a trend for patients with NPWT treated fractures to be discharged sooner, 2.5 days compared to 3.0 days. NPWT delivers negative pressure suction through a closed system beneath a sealed adhesive film to promote wound healing through multiple mechanisms of action.

With respect to the measure specifications, we support the numerator and denominator statements, but would suggest that the rationale should include both a reference to the 2012 study Stannard JP et al. Incisional Negative Pressure Wound Therapy After High-Risk Lower Extremity Fractures. *J Orthop Trauma*2012 Jan; 26(1):37-42, and specific reference to treatments such as NPWT that have been shown to lower the risk of infections for patients experiencing tibia plateau fractures.

In sum, we believe this measure would advance patient care and we urge the NQF to endorse this measure.

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 43T

Measure Title: Infection rate in Bicondylar tibia plateau fractures

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 43T

Date of Submission: 6/8/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☒ Health outcome: Infection rate in bicondylar tibia plateau fracture

☐Patient-reported outcome (PRO): 43T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 43T

☐ Process: 43T

☐ Structure: 43T

☐ Other: 43T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Bicondylar tibial plateau are difficult injuries to treat and are often complicated by infection, nonunion, and compartment syndrome. Infection rate of these injuries is reported to be between 20-30% at high volume centers with experienced surgeons following a staged fixation protocol with dual incisions. The lowest infection rate ever reported for bicondylar tibia plateau fractures treated with ORIF is 8%. As orthopedic surgeons the OTA hopes that we can "do better than this" Routine reporting of infection rates associated with these injuries can help all surgeons evaluate their results and lead to improved Quality Improvement processes to drive down the current infection rates. These injuries were chosen because they are commonly treated by orthopedic surgeons with advanced Trauma training and they have some of the highest reported rates of this infection of any operation. Infection increases the cost of care as treatment involves multiple surgical debridements, multiple clinic and hospital visits and often increases the length of stay during inpatient care. This long treatment course also immobilizes the patient which affects their physical and emotional health.

Orthopedic surgeons cannot change the injury or patient, but can attempt to optimize modifiable patient characteristics, control timing and type of fixation, surgical technique and post op protocols for treatment of these injuries. No one action will likely be detectable, but a series of actions and processes that are in control of the surgeon can be implemented that may decrease the high infection rates that are reported. Surgeons with high rates can re-evaluate their actions and processes and surgeons with low rates can publish their processes to encourage best practices by our colleagues.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

A deep wound infection increases the morbidity of the injury and treatment on the patient. Infection raises the risk of nonunion and certainly increases number of surgeries required, pain suffered and intensity of interventions.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Healing a severe injury like a bicondylar treatment tibia plateau fracture without a skin muscle or deep infection intuitively has significant improvements in patient health and activity.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.

Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.

Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.

Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.

Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

Fakler, Johannes KM, et al. "Optimizing the management of Moore type I postero-medial split fracture dislocations of the tibial head: description of the Lobenhoffer approach." Journal of orthopaedic trauma 21.5 (2007): 330-336.

Stamer, David T., et al. "Bicondylar tibial plateau fractures treated with a hybrid ring external fixator: a preliminary study." Journal of orthopaedic trauma 8.6 (1994): 455-461.

Partenheimer, A., et al. "[Management of bicondylar fractures of the tibial plateau with unilateral fixed-angle plate fixation]." Der Unfallchirurg 110.8 (2007): 675-683.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____ Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.

Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.

Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.

Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.

Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

Fakler, Johannes KM, et al. "Optimizing the management of Moore type I postero-medial split fracture dislocations of the tibial head: description of the Lobenhoffer approach." Journal of orthopaedic trauma 21.5 (2007): 330-336.

Stamer, David T., et al. "Bicondylar tibial plateau fractures treated with a hybrid ring external fixator: a preliminary study." Journal of orthopaedic trauma 8.6 (1994): 455-461.

Partenheimer, A., et al. "[Management of bicondylar fractures of the tibial plateau with unilateral fixed-angle plate fixation]." Der Unfallchirurg 110.8 (2007): 675-683.

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review? Fracture healing without a deep infection is the outcome measure.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 43T

N/A

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

3 retrospective observarional studies

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Outcomes are improved if patient do not obtain a deep wound infection after surgery of a severe proximal tibia fracture. Infection delays soft tissue healing and potentially bone healing.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Potential harms are deep infection, nonunion, or amputation. Avoiding any of these outcome measures is associated with significant improvement to the patient.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

Literature review of the current available data

1a.8.2. Provide the citation and summary for each piece of evidence.

Bicondylar tibial plateau are difficult injuries to treat and are often complicated by infection, nonunion, and compartment syndrome. Infection rate of these injuries is reported to be between 20-30% at high volume centers with experienced surgeons following a staged fixation protocol with dual incisions. The lowest infection rate ever reported for bicondylar tibia plateau fractures treated with ORIF is 8%. As orthopedic surgeons the OTA hopes that we can "do better than this" Routine reporting of infection rates associated with these injuries can help all surgeons evaluate their results and lead to improved Quality Improvement processes to drive down the current infection rates. These injuries were chosen because they are commonly treated by orthopedic surgeons with advanced Trauma training and they have some of the highest reported rates of this infection of any operation. Infection increases the cost of care as treatment involves multiple surgical debridements, multiple clinic and hospital visits and often increases the length of stay during inpatient care. This long treatment course also immobilizes the patient which affects their physical and emotional health. Orthopedic surgeons cannot change the injury or patient, but can attempt to optimize modifiable patient characteristics, control timing and type of fixation, surgical technique and post op protocols for treatment of

these injuries. No one action will likely be detectable, but a series of actions and processes that are in control of the surgeon can be implemented that may decrease the high infection rates that are reported. Surgeons with high rates can re-evaluate their actions and processes and surgeons with low rates can publish their processes to encourage best practices by our colleagues.

A deep wound infection increases the morbidity of the injury and treatment on the patient. Infection raises the risk of nonunion and certainly increases number of surgeries required, pain suffered and intensity of interventions.

The literature has consistently shown a high rate of deep infection when treating bicondylar tibial plateau fractures. Barei et al 8 had an 8.4%(7/83) infection rate. Ruffolo et al. retrospectively analyzing 140 bicondylar tibial plateau fractures had an infection rate of 23.6 (33/140). Morris et al through a retrospective review on 302 fractures had a 14.2%(43/302) infection rate.

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

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

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

evidence_attachment_Bicondylar_tibia_plateau_fxs_6-8-16_-1-.docx

1b. Performance Gap

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

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

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

Bicondylar tibial plateau are difficult injuries to treat and are often complicated by infection, nonunion, and compartment syndrome. Infection rate of these injuries is reported to be between 20-30% at high volume centers with experienced surgeons following a staged fixation protocol with dual incisions. The lowest infection rate ever reported for bicondylar tibia plateau fractures treated with ORIF is 8%. As orthopedic surgeons the OTA hopes that we can "do better than this" Routine reporting of infection rates associated with these injuries can help all surgeons evaluate their results and lead to improved Quality Improvement processes to drive down the current infection rates. These injuries were chosen because they are commonly treated by orthopedic surgeons with advanced Trauma training and they have some of the highest reported rates of this infection of any operation. Infection increases the cost of care as treatment involves multiple debridements, multiple clinic and hospital visits and often increases the length of stay during inpatient care. This long treatment course also immobilizes the patient which affects their physical and emotional health. Orthopedic surgeons cannot change the injury or patient, but can attempt to optimize modifiable patient characteristics, control timing and type of fixation , surgical technique and post op protocols for treatment of these injuries. No one action will likely be detectable, but a series of actions and processes that are in control of the surgeon can be implemented that may decrease the high infection rates that are reported. Surgeons with high rates can re-evaluate their actions and processes and surgeons with low rates can publish their processes to encourage best practices by our colleagues.

A deep wound infection increases the morbiity of the injury and treatment on the patient. Infection raises the risk of nonunion and certainly increases number of surgeries required, pain suffered and intensity of interventions.

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.

Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.

Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.

Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.

Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

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

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
The literature has consistently shown a high rate of deep infection when treating bicondylar tibial plateau fractures. Barei et al had an 8.4%(7/83) infection rate. Ruffolo et al. retrospectively analyzing 140 bicondylar tibial plateau fractures had an infection rate of 23.6 % (33/140). Morris et al through a retrospective review on 302 fractures had a 14.2%(43/302) infection rate.

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.
Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.
Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.
Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.
Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.
Fakler, Johannes KM, et al. "Optimizing the management of Moore type I postero-medial split fracture dislocations of the tibial head: description of the Lobenhoffer approach." Journal of orthopaedic trauma 21.5 (2007): 330-336.
Stamer, David T., et al. "Bicondylar tibial plateau fractures treated with a hybrid ring external fixator: a preliminary study." Journal of orthopaedic trauma 8.6 (1994): 455-461.
Partenheimer, A., et al. "[Management of bicondylar fractures of the tibial plateau with unilateral fixed-angle plate fixation]." Der Unfallchirurg 110.8 (2007): 675-683.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Not available. We can include SES and demographic data in the QCDR to be able to analyze at a later date

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
The literature has consistently shown a high rate of deep infection when treating bicondylar tibial plateau fractures. Barei et al had an 8.4%(7/83) infection rate. Ruffolo et al. retrospectively analyzing 140 bicondylar tibial plateau fractures had an infection rate of 23.6 (33/140). Morris et al through a retrospective review on 302 fractures had a 14.2%(43/302) infection rate.

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.
Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.
Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.
Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.
Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.
Fakler, Johannes KM, et al. "Optimizing the management of Moore type I postero-medial split fracture dislocations of the tibial head: description of the Lobenhoffer approach." Journal of orthopaedic trauma 21.5 (2007): 330-336.
Stamer, David T., et al. "Bicondylar tibial plateau fractures treated with a hybrid ring external fixator: a preliminary study." Journal of orthopaedic trauma 8.6 (1994): 455-461.
Partenheimer, A., et al. "[Management of bicondylar fractures of the tibial plateau with unilateral fixed-angle plate fixation]." Der Unfallchirurg 110.8 (2007): 675-683.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

A leading cause of morbidity/mortality, Patient/societal consequences of poor quality, Severity of illness

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

The literature has consistently shown a high rate of deep infection when treating bicondylar tibial plateau fractures. Barei et al had an 8.4%(7/83) infection rate. Ruffolo et al. retrospectively analyzing 140 bicondylar tibial plateau fractures had an infection rate of 23.6 (33/140). Morris et al through a retrospective review on 302 fractures had a 14.2%(43/302) infection rate.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Basques, Bryce A., et al. "Adverse events, length of stay, and readmission after surgery for tibial plateau fractures." Journal of orthopaedic trauma 29.3 (2015): e121-e126.

Barei, David P., et al. "Complications associated with internal fixation of high-energy bicondylar tibial plateau fractures utilizing a two-incision technique." Journal of orthopaedic trauma 18.10 (2004): 649-657.

Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.

Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.

Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

Fakler, Johannes KM, et al. "Optimizing the management of Moore type I postero-medial split fracture dislocations of the tibial head: description of the Lobenhoffer approach." Journal of orthopaedic trauma 21.5 (2007): 330-336

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Prevention, Safety : Complications, Safety : Healthcare Associated Infections

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

N/A

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

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

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients aged 18 years and older undergoing ORIF of a bicondylar tibial plateu fracture who develop a postoperative deep incisional infection associated with an implant within 1 year of fracture fixation. We do not have adequate data to provide adequate risk stratification at this time.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
3 years

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Deep incisional SSI Must meet the following criteria:
Infection occurs within 1 year after the index operative procedure (where day 1 = the procedure date)
AND
involves deep soft tissues of the incision (e.g., fascial and muscle layers)
AND
patient has at least one of the following: a. purulent drainage from the deep incision. b. a deep incision that spontaneously dehisces, or is deliberately opened or aspirated by a surgeon, attending physician** or other designee and an
organism is identified by a culture or non-culture based microbiologic testing method which is performed for purposes of clinical diagnosis or treatment (e.g., not Active January 2016 9-9 Procedure-associated Module SSI Surveillance Culture/Testing (ASC/AST) or culture or non-culture based microbiologic testing method is not performed
AND
patient has at least one of the following signs or symptoms: fever (>38°C); localized pain or tenderness. A culture or non-culture based test that has a negative finding does not meet this criterion.

Through patient records, patients with closed bicondylar tibial plateau fractures will be identified. Patients for this study will be selected by narrowing down the pool of patients with those who have the complication of deep infection.
Patient with infection will be identified by an operative report for irrigation and debridement of the operative wound and confirmed culture-positive intraoperative findings. Patients can be identified with either and ICD-10 code (S82.141, S82.142) or by CPT billing codes. (27536) and have an admission for a post op wound infection (CPT 10180)

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
All patients undergoing ORIF of a closed bicondylar tibial plateau fracture aged 18 years or older. Patients can be identified with either and ICD-10 code (S82.141, S82.142) or by CPT billing codes. (27536). Risk calculation can be added once adequate volume of patients are enrolled.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Number of bicondylar tibial plateau procedures utilizing ICD-10 codes S82.141 (right tibia) and S82.142 (left tibia) and have a procedure for fixation of this injury with CPT code 27536 utilized

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
N/A

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
N/A

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
We are not able to perform risk stratification at this time. We will gather the data below as well as previously reported risk factors for infection in the orthopedic literature for this injury.Previously reported factors in relatively small case series associated with infection in closed bicondylar tibia plateaus are: open fractures, fasciotomy wounds, smoking, compartment syndrome, 2 incisions for fixation., dysvascular limbs, ASA score > 3, male sex, and pulmonary disease.

Other factors associated with increased risk for any SSI as judged by CDC are:

ASA physical status: Assessment by the anesthesiologist of the patient's preoperative physical condition using the American Society of Anesthesiologists' (ASA) Classification of Physical Status12,13 . Patient is assigned one of the following: 1. A normally healthy patient 2. A patient with mild systemic disease 3. A patient with severe systemic disease 4. A patient with severe systemic disease that is a constant threat to life 5. A moribund patient who is not expected to survive without the operation.
Diabetes: The NHSN SSI surveillance definition of diabetes indicates that the patient has a diagnosis of diabetes requiring management with insulin or a non-insulin anti-diabetic agent. This includes patients with "insulin resistance" who are on management with anti-diabetic agents. This also includes patients with a diagnosis of diabetes who are noncompliant with their diabetes medications.
Duration of operative procedure: The interval in hours and minutes between the Procedure/Surgery Start Time, and the Procedure/Surgery Finish Time, as defined by the Association of Anesthesia Clinical Directors (AACD) 14 : • Procedure/Surgery Start Time (PST): Time when the procedure is begun (e.g., incision for a surgical procedure). • Procedure/Surgery Finish (PF): Time when all instrument and sponge counts are completed and verified as correct, all postoperative radiologic studies to be done in the OR are completed, all dressings and drains are secured, and the physicians/surgeons have completed all procedure-related activities on the patient. Emergency operative procedure: A nonelective, unscheduled operative procedure.

Emergency operative procedures are those that do not allow for the standard immediate preoperative preparation normally done within the facility for a scheduled operation (e.g., stable vital signs, adequate antiseptic skin preparation, etc.).
General anesthesia: The administration of drugs or gases that enter the general circulation and affect the central nervous system to render the patient pain free, amnesic, unconscious, and often paralyzed with relaxed muscles. This does not include conscious sedation. Height:
Non-primary Closure is defined as closure that is other than primary and includes surgeries in which the skin level is left completely open during the original surgery and therefore cannot be classified as having primary closure. For surgeries with non-primary closure, the deep tissue layers may be closed by some means (with the skin level left open), or the deep and superficial layers may both be left completely open. An example of a surgery with non-primary closure would be a laparotomy in which the incision was closed to the level of the deep tissue layers, sometimes called "fascial layers" or "deep fascia," but the skin level was left open. Another example would be an "open abdomen" case in which the abdomen is left completely open after the surgery. Wounds with non-primary closure may or may not be described as "packed" with January 2016 9-5 Procedure-associated Module SSI gauze or other material, and may or may not be covered with plastic, "wound vacs," or other synthetic devices or materials. Primary Closure is defined as closure of the skin level during the original surgery, regardless of the presence of wires, wicks, drains, or other devices

or objects extruding through the incision. This category includes surgeries where the skin is closed by some means. Thus, if any portion of the incision is closed at the skin level, by any manner, a designation of primary closure should be assigned to the surgery. . Trauma: Blunt or penetrating injury occurring prior to the start of the procedure.
Weight: The patient's most recent weight documented in the medical record in pounds (lbs.) or kilograms (kg) prior to or otherwise closest to the procedure.
Wound class: An assessment of the degree of contamination of a surgical wound at the time of the operation. Wound class should be assigned by a person involved in the surgical procedure (e.g., surgeon, circulating nurse, etc.). The wound class system used in NHSN is an adaptation of the American College of Surgeons wound classification schema. Wounds are divided into four classes: 1. Clean: An uninfected operative wound in which no inflammation is encountered and the respiratory, alimentary, genital, or uninfected urinary tracts are not entered. In addition, clean wounds are primarily closed and, if necessary, drained with closed drainage. Operative incisional wounds that follow nonpenetrating (blunt) trauma should be included in this category if they meet the criteria. Note: The clean wound classification level will not be available for denominator data entry for the following NHSN operative procedure categories: APPY, BILI, CHOL, COLO, REC, SB, and VHYS 2. Clean-Contaminated: Operative wounds in which the respiratory, alimentary, genital, or urinary tracts are entered under controlled conditions and without unusual contamination. Specifically, operations involving the biliary tract, appendix, vagina, and oropharynx are included in this category, provided no evidence of infection or major break in technique is encountered. 3. Contaminated: Open, fresh, accidental wounds. In addition, operations with major breaks in sterile technique (e.g., open cardiac massage) or gross spillage from the gastrointestinal tract, and incisions in which acute, non-purulent inflammation is encountered including necrotic tissue without evidence of purulent drainage (e.g., dry gangrene) are included in this category. 4. Dirty or Infected: Includes old traumatic wounds with retained devitalized tissue and those that involve existing clinical infection or perforated viscera.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF a PRO-PM, identify whether (and how) proxy responses are allowed.
All cases that qualify will be included in the denominator and numerator

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry, Other

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
An OTA certified QCDR will be used by OTA members to gather and record data elements and outcomes. The OTA will publish data elements and outcome measure on public web site so non-OTA members are able to keep their own database using this Performance Measure.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Clinician : Individual, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
testing_attachmentBicondylaPlatea_6-8_16.docx,ICD_and_CPT_codes.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 43T
Measure Title: Infection rate of bicondylar tibia plateau fractures43T
Date of Submission: 6/9/2016
Type of Measure:

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

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set***

of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.

- **For <u>all</u> measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For <u>outcome and resource use</u> measures**, section **2b4** also must be completed.
- If specified for **<u>multiple data sources/sets of specificaitons</u>** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to <u>all</u> questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

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

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

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

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

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

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

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

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

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

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims

☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 43T	☐ other: 43T

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

1.3. What are the dates of the data used in testing? 43T

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

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 43T	☐ other: 43T

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*) Data was pooled from two institutions that included data from orthopedic trauma surgeons (10). This group of surgeons accurately represent the type of clinicians who commonly treat bicondylar tibial plateau fractures

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)
All 302 patients in the study underwent open reduction and internal fixation. 44/302 patients in the study developed deep infection and had a subsequent admission for a surgical site infection at the index site with an infection based on CDC criteria. 193(63.9%) of patients were male with a mean age of 45.7 years. The mechanism of injury was motor vehicle crash (54.3%), fall from height (18.9%), motorcycle crash (10.6%), pedestrian versus car crash (7.3%), and others (8.9%).

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

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy

variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).
Socioeconomic status, patient factors and injury factors have not consistently been associated with differences in surgical site infection rates of patients with bicondylar tibia plateau fractures that underwent internal fixation. All patients with this injury that undergo surgical stabilization will be evaluated to determine the rate of deep infection in all patients treated.

We propose an evaluation and transition period. We would like to gather sociodemographic status (SDS) and clinical factors for a period of 3 year with the data being available to the Orthopedic Trauma Association during this time.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☐ **Performance measure score** (e.g., *signal-to-noise analysis*) CDC criteria for deep infection will be utilized as a national standard for definition of deep infection and injury is defined as a bicondylar tibia plateau fracture (ICD-10 S82.14) and treated with internal fixation (CPT 27536).

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

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

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

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

- ☐ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
All 77 fractures with the diagnosis of infected bicondylar tibial plateau fractures From the Morris et al and the Ruffolo et al articles were confirmed to be bicondylar tibial plateau fractures through the use of radiographs and CT scan by an orthopedic specialist. 76/77 of these bicondylar tibial plateau fractures were confirmed to be infected through an operative report for irrigation and debridement, and organism positive laboratory data one

patient who was thought to have infection was determined to have a seroma which is a noninfectious build up of fluid at the site of the incision. From patients enrolled in the Morris et al and the Ruffolo et al articles a random sample of uninfected patients demonstrated

	A	B	Total
A	76	1	77
B	0	95	95
Total	76	96	172

Number of observed agreements: 171 (99.42% of the observations)
Number of agreements expected by chance: 87.0 (50.61% of the observations)

Kappa= 0.988
SE of kappa = 0.012
95% confidence interval: From 0.965 to 1.000
The strength of agreement is considered to be 'very good'.

95/95 bicondylar tibial plateau fractures with no infection were confirmed to have bicondylar tibial plateau fractures through radiographs and CT scans. 95/95 bicondylar tibial plateau fractures with no diagnosis of infection were confirmed to have NO deep wound infection based on the lack of presence of operative reports for irrigation and debridement and no laboratory data indicating the presence of infection.

	Infected	Noninfected	Total
Positive	76	1	T positive= 77
Negative	0	95	T negative=95
	76	96	Total=172

Sensitivity=100%
Specificity=99%
Positive Predictive Value=98.7%
Negative Predictive Value= 100%

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
Results

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*). It is feasible to get the results from registry and claims data. This data can be pulled from medical records and there is no foreseeable reason that the information could not be accrued from a Registry. As we accrue the patients by evaluating CPT data, we have no reason to believe that the same information cannot be reliably extracted from claims data and an EMR_____

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance*

measure scores)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with 43T risk factors**
☐ **Stratification by 43T risk categories**
☐ **Other,** 43T

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.
Factors that have been identified in the orthopedic literature to be associated with deep infection include a dysvascular limb, pulmonary disease, male sex, smoking, ASA score greater than or equal to 3 open fractures, fasciotomies that have not been closed.
The incidence of deep infection after fixation of bicondylar tibia plateau fractures has not been associated with variation in social economic or injury patterns.
Of the 43 patients with a deep infection from one institution (Morris et al) , the average age was 46.9 years, 76.7% were male, 83.7% were white,62.8% reported smoking, and 7.0% had a diagnosis of diabetes. Thirty-one of the 43 patients had a 2-incision technique for surgical fracture fixation with 2 plates (72.1%). These variables (dysvascular limb, pulmonary disease, male sex, smoking, ASA score greater than or equal to 3 open fractures, fasciotomies that have not been closed) are known risk factors for infection in high energy trauma.
Our patient population accurately represents the demographic of patients who develops infection. We have no reason to believe that the demographics would be different in other institutions.
The table below compares the demographics of the 4 most cited single center series of bicondylar tibia plateau fractures and infection rates.

	Morris	Ruffolo	Basques	Barei
Age	45.7±14.3 yrs	44.6 yrs	52.3±15.8 yrs	N/A
Race	85.1% white 14.9% non-white	N/A	N/A	N/A
Gender	63.9% Male 36.1% Female	68% Male 32% Female	42.8% Male 57.2% Female	N/A
Open/closed Fx	16.6% open fractures 83.4% closed fractures	11.4% Open Fractures 88.6% Closed	1.9 % Open 98.1% Closed	Open 13.3% Closed 86.7%
Fracture Type	AO/OTA 41-C	AO/OTA 41-C3	Unicondylar 70.9% Bicondylar 29.1%	N/A

Compartment Syndrome	Yes 7.3% No 92.7%	Yes 17.9% No 82.1%	Yes 1.2% No 98.8%	Yes 14.5%
Time to Fixation	10.6±10.6 Days	12.5 Days	N/A	9 Days
Tobacco Use	Yes 45.4% No 54.6%	38.6% Yes 61.4% No	Yes 32.8% No 67.2%	N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)
Per the data in the table above, there is minimal variation between facilities.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE
2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified

(*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*
Differences in deep infection rate have varied from 8% to 23% reported at the institutional level. This variation indicates that improvement at the institutional and individual level are feasible. We do not have enough sample size to detect differences at surgeon level.

- After gathering demographic and clinically relevant data for a 3 year time, we would then perform univariate analysis on each factor to determine if it is potentially a predictor of deep infection. We would then include all variables that were statistically significant in univariate analysis and perform a multivariate analysis. By identifying the risk factors of open fractures types and smoking, we hope to reduce the number of infected bicondylar tibial plateaus. In our study 27(62.8%) of patients with infection were smokers (P=.82) and 13 (30.2%) of patients with infection had a history of open fracture (P=.01)

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.
N/A – only source is Registry

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Database will be built so that all data fields need to be completed.

No known missing data in Testing

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

In a study by Morris et al., 14.2% (43/302) were complicated with infection that required reoperation. 46.5% (20/43) of the infections were MRSA positive. 23.6% of the bicondylar tibial plateau fractures in the study by Ruffolo et al were complicated by infection as well.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

ICD-10 will provide specificity of injury and diagnoses that ration of injuries with and without infection will be able to be calculated.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
This performance measure is feasible to obtain and perform. This measure will require high compliance and utilization by individual surgeons and systems. When collecting and verifying the data, few inconsistencies were found in the operative reports, labs and fracture classifications in a secondary evaluation of the two largest series reported in the orthopedic literature. To assess data quality, we reviewed the medical charts of the patient's identified to have infection from the 2 largest series of closed bicondylar tibial plateau fractures in the literature (see references below) 76/77 of the patient's that were rechecked and identified to have a deep infection actually had culture-positive results from an operative debridement. One patient had a debridement of a fluid collection with a negative culture. 95 patients randomly selected from these 2 studies were again evaluated and all 95/95 had a closed bicondylar tibia plateau fracture on xray and did NOT have any evidence of deep infection. This indicates that data collection is feasible and accurate.

Ruffolo, Michael R., et al. "Complications of High-Energy Bicondylar Tibial Plateau Fractures Treated With Dual Plating Through 2 Incisions." Journal of orthopaedic trauma 29.2 (2015): 85-90.
Morris, Brent J., et al. "Risk factors of infection after ORIF of bicondylar tibial plateau fractures." Journal of orthopaedic trauma 27.9 (2013): e196-e200.
Ahearn, N., et al. "The outcome following fixation of bicondylar tibial plateau fractures." Bone & Joint Journal 96.7 (2014): 956-962.

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

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Planned	Current Use (for current use provide URL)
Public Reporting	

Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

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

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
The Orthopedic Trauma Association has not had quality metrics for public reporting. This measure could also be used by the American College of Surgeons as a quality indicator in their review of trauma centers. The OTA currently has an approved QCDR. The QCDR is not currently transparent to public, but could be made to be transparent to the OTA membership and public. The OTA could also publish the overall infection rates and ranges so any surgeon or center could internally review their own cases and surgeons

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

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of**

unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.
No unintended negative consequences to individuals or populations were identified during testing

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
Mo other performance measure is similar or for same target population.

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

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

Attachment **Attachment:** OTA_BOD_signed_agreement.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Orthopedic Trauma Association
Co.2 **Point of Contact:** William, Obremskey, william.obremskey@vanderbilt.edu, 615-260-2054-
Co.3 **Measure Developer if different from Measure Steward:** Orthopedic Trauma Association
Co.4 **Point of Contact:** William, Obremskey, william.obremskey@vanderbilt.edu, 615-260-2054-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
Jaimo Ahn -Jaimo.Ahn@uphs.upenn.edu>
Arvind Anna - ´anana@jpshealth.org
Chad Coles - coles@dal.ca
Cory Collinge- cory.a.collinge@vanderbilt.edu>;
Gundrum Mirick - gmirick@gmail.com
Michael Zlowodzki - zlowi@web.de
Steve Olson - olson016@mc.duke.edu
Paul Tornetta (ptornetta@gmail.com
Role - All members of the OTA EBQVS Committee review, revise and approve Measure

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**

CPT Code

1. CPT code 27536-

Open treatment of tibial fracture, proximal(plateau): bicondylar, with or without internal fixation

ICD-10 Codes Open Fractures

S82.142B Displaced bicondylar fracture of left tibia, initial encounter for open fracture type I or II

...Displaced bicondylar fracture of left tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.142C Displaced bicondylar fracture of left tibia, initial encounter for open fracture type IIIA, IIIB, or IIIC

...Displaced bicondylar fracture of left tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.142E Displaced bicondylar fracture of left tibia, subsequent encounter for open fracture type I or II with routine healing

...Displaced bicondylar fracture of left tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.142F Displaced bicondylar fracture of left tibia, subsequent encounter for open fracture type IIIA, IIIB, or IIIC with routine healing

...Displaced bicondylar fracture of left tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.141B Displaced bicondylar fracture of right tibia, initial encounter for open fracture type I or II

...Displaced bicondylar fracture of right tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.141C Displaced bicondylar fracture of right tibia, initial encounter for open fracture type IIIA, IIIB, or IIIC

...Displaced bicondylar fracture of right tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.141E Displaced bicondylar fracture of right tibia, subsequent encounter for open fracture type I or II with routine healing

...Displaced bicondylar fracture of right tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.141F Displaced bicondylar fracture of right tibia, subsequent encounter for open fracture type IIIA, IIIB, or IIIC with routine healing

...Displaced bicondylar fracture of right tibia,...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

ICD-10 Codes Closed Fractures

S82.141A Displaced bicondylar fracture of right tibia, initial encounter for closed fracture

...tibia, initial encounter for closed fracture — Fracture of proximal endof proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.141D Displaced bicondylar fracture of right tibia, subsequent encounter for closed fracture with routine healing

...tibia, subsequent encounter for closed fracture with routine healing — Fracture...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.142 Displaced bicondylar fracture of left tibia S82.142A Displaced bicondylar fracture of left tibia, initial encounter for closed fracture

...tibia, initial encounter for closed fracture — Fracture of proximal end of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

S82.142D Displaced bicondylar fracture of left tibia, subsequent encounter for closed fracture with routine healing

...tibia, subsequent encounter for closed fracture with routine healing — Fracture...of proximal end of tibia — Fracture of tibial plateau NOS — fracture...of upper end of tibia — Bicondylar fracture of tibia — Displaced bicondylar...

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3016
De.2. **Measure Title:** PBM-01: Preoperative Anemia Screening
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure assesses the proportion of selected elective surgical patients age 18 and over with documentation of pre-operative anemia screening in the window between 45 and 14 days before the surgery start date
1b.1. **Developer Rationale:** Researchers have shown that preoperative hemoglobin (hgb) and hematocrit can be used as predictors of outcome for specific types of patients such as cardiac artery bypass graft or orthopedic surgery. 1 Preoperative anemia and exposure to allogeneic transfusion are associated with increased morbidity and mortality after surgery4. Previously undiagnosed anemia was identified in 5% - 75% of elective surgery patients in certain populations and a national audit demonstrated that 35% of patients scheduled for joint replacement therapy have a hgb. < 13 g/dL on preadmission testing.2 In the elderly (>65 yr. old), the prevalence of anemia as defined by the World Health Organization (WHO) is 11% and 10.2% for men and women, respectively.3 Blood transfusions are associated with several postsurgical complications, including surgical site infections, pneumonia, slower wound healing, prolonged ventilator use and increased length of stay.4

Development of formal protocols for preoperative testing of Hgb for potentially high blood loss elective surgeries could be used to identify and intervene for optimal management of blood resources. Early recognition of anemia offers patients an opportunity to receive the most appropriate transfusion-sparing strategy and avoid the risk of a potential transfusion. A panel of multidisciplinary physicians developed a clinical care pathway for anemia management in the elective surgical patients for whom blood transfusion is a probability (defined as any procedure for which a preoperative blood type and crossmatch is requested). They recommend that "Whenever clinically feasible, elective surgical patients should have a Hgb level tested a minimum of 30 days before the scheduled surgical procedure".2 Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need.4

1. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
2. Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005;101:1858-61.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. Br. Journ. Anesthesia, 106 (1): 13-22 (2011).

S.4. **Numerator Statement:** Patients with preoperative anemia screening done in the window between 45 and 14 days prior to the surgery start date.
S.7. **Denominator Statement:** Patients age 18 and older with a length of stay less than or equal to 120 days who undergo selected elective surgical procedures
S.10. **Denominator Exclusions:** • Patients whose surgical procedure is performed to address a traumatic injury • * Patients with a solid organ transplant recorded <=48 hours prior to the encounter or during the encounter

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

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

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

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

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

Evidence summary

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**
-
- The developer stated the path to support the relationship between the process of timely assessment for anemia prior to elective surgery and the outcome of a reduced risk of transfusion-related adverse outcomes.
- The developer provided a guideline from The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists:
    - Preoperative identification of high-risk patients (advanced age, preoperative anemia, small body size, noncoronary artery bypass graft or urgent operation, preoperative antithrombotic drugs, acquired or congenital coagulation/clotting abnormalities and multiple patient comorbidities) should be performed, and all available preoperative and perioperative measures of blood conservation should be undertaken in this group as they account for the majority of blood products transfused. **Class I: Level of Evidence B**
- The developer presented two systematic reviews that addressed the detection and treatment of preoperative anemia. Summaries of the Quantity, Quality, and Consistency (QQC) of the body of evidence were partially provided.
- The developer provided an additional four citations as a source of evidence for this measure.
- The developer did not provide specific evidence to support the 14-45 day prior to surgery timeframe for preoperative anemia screening.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

Based on SR/grading for cardiac surgery (Box 3) → QQC partially provided for 2 systematic reviews relevant to general and orthopedic surgery (Box 4) → Moderate quality evidence (Box 5b) → MODERATE

Questions for the Committee:

- *Is the evidence directly applicable to the process of timely assessment for anemia prior to elective surgery and the outcome of a reduced risk of transfusion-related adverse outcomes?*
    - *How strong is the evidence for this relationship?*

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

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

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

- Data from the literature demonstrating a performance gap was not provided. This is a new measure and as testing has not yet be done, gap data from testing was not provided.
- The developer provided the following unpublished data from a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015: "*Respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 118 of the 141 respondents (81%) indicated that there was a gap in practice; 6 were not sure, and 17 reported no gap. Of the 118, most indicated that pre-operative anemia screening was done 3 or 4 days in advance of the elective surgical procedure*".

Disparities

- The developed indicated that no disparity data are available.

Questions for the Committee:

- *Do you agree that the data provided by the developer demonstrates that there is a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1a. Evidence to Support the Measure Focus

- Depends on which surgeries are included, which might be a moving target Unnecessary pre-op testing is a major burden for healthcare.
  Feasibility difficulty with tests done outside the system.
  OK with trialing measure

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients with preoperative anemia screening done in the window between 45 and 14 days prior to the surgery start date.
- Denominator Statement: Patients age 18 and older with a length of stay less than or equal to 120 days who undergo selected elective surgical procedures
- Denominator Exclusions:
  - Patients whose surgical procedure is performed to address a traumatic injury
  - Patients with a solid organ transplant recorded <=48 hours prior to the encounter or during the encounter
- Level of Analysis: Facility
- Care Setting: Hospital/Acute Care Facility

- No risk adjustment or risk stratification

eMeasure Technical Advisor(s) review

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously; - Submitted with results from Bonnie testing
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

Questions for the Committee :

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

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

- Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 23 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications are consistent with evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.48
Denominator clearly describes the activity being measured	4.48
Numerator inclusions clear and appropriate	4.48
Denominator inclusions clear and appropriate	4.48
Numerator exclusions clear and appropriate	4.48
Denominator exclusions clear and appropriate	4.48
Accurately assesses the process of care to which it is addressed	4.28

This measure is being considered for trial use, thus full validity testing results are not expected.

2b3-2b7. Threats to Validity

2b3. Exclusions:
When data are available, the developer plans to analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid organ transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury

Questions for the Committee:

- *Are there other threats to validity the measure developer should consider?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time.

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

- The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

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

Committee pre-evaluation comments

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

Criterion 3. Feasibility

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

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

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

Publicly reported? ☐ **Yes** ☒ **No**

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

Accountability program details The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms None identified

Feedback : None indicated

Questions for the Committee*:*

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

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

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 43T

Measure Title: PBM-01: Preoperative Anemia Screening

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 43T

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*

- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☐ Health outcome: 43T

☐Patient-reported outcome (PRO): 43T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 43T

☒ Process: Timely assessment for anemia prior to elective surgery

☐ Structure: 43T

☐ Other: 43T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1. *Process: Timely assessment for anemia prior to elective surgery*
2. *Investigation into cause for anemia*
3. *Correction of anemia prior to surgery*
4. *Reduced perioperative transfusion rate*
5. *Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which are decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.*

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Ferraris VA, Brown JR, Despotis GJ, Hammon JW, Reece TB, Saha SP, Song HK, Clough ER, Shore-Lesserson LJ, Goodnough LT, Mazer CD, Shander A, Stafford-Smith M, Waters J, Baker RA, Dickinson TA, Fitzgerald DJ, Likosky DS, Shann KG. 2011 update to the Society of Thoracic Surgeons and the Society of Cardiovascular Anesthesiologists blood conservation clinical practice guidelines. *Ann Thorac Surg*. 2011 Mar;91(3):944-82. [404 references.

http://www.annalsthoracicsurgery.org/article/S0003-4975(10)02888-2/abstract

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Class I

Preoperative identification of high-risk patients (advanced age, preoperative anemia, small body size, noncoronary artery bypass graft or urgent operation, preoperative antithrombotic drugs, acquired or congenital coagulation/clotting abnormalities and multiple patient comorbidities) should be performed, and all available preoperative and perioperative measures of blood conservation should be undertaken in this group as they account for the majority of blood products transfused. (Level of evidence A)

Guideline citation is: Ferraris V, Ferraris S, Saha SP, Hessel EA, et al. Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline. *Ann Thorac Sur*g 2007;83:S27-86, page S31. This guideline remained unchanged in the updated Guideline cited above in 1a.4.1.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

CLASS I

Benefit >>> Risk

Procedure/Treatment
SHOULD be performed/administered

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

CLASS IIa	**CLASS IIb**	**CLASS III**
Benefit >> Risk *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	*Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	*Risk ≥ Benefit* Procedure/Treatment should **NOT** be performed/administered **SINCE IT IS NOT HELPFUL AND MAY BE HARMFUL**

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines

http://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/documents/downloadable/ucm_319826.pdf

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

REVIEW #1:

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. *Br. Journ. Anesthesia*, 106 (1): 13-22 (2011).

http://bja.oxfordjournals.org/content/106/1/13.full

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Same as 1a.6.1

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Detection of anemia.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

We recommend that elective surgical patients have an Hb level determination as close to 28 days before the scheduled surgical procedure as possible. Grade 1C

Grade 1: Risk/benefit is clear, strong recommendation

Grade C: Observational studies, RCTs with major limitations

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Grade 1: Weak recommendation, risk/benefit not clear

Grade A: Meta-analysis, RCTs

Grade B: RCTs with limitations, observational studies with large effects

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1966 – January 2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*) Specific studies supporting this recommendation were not enumerated.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not specifically stated

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Grade is 1, indicating that benefit exceeds any risk

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Unstated

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

None

REVIEW #2:

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Kotze A, Harris A, Baker C, Iqbal T, eta l. British Committee for Standards in Haemotology Guidelines on the Identification and Management of Pre-Operative Anemia. *British Journal of Haemotology* Volume 171, Issue 3: November 2015 pages 322-331.

http://onlinelibrary.wiley.com/enhanced/doi/10.1111/bjh.13623/

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Same as 1a.6.1

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Investigation and treatment of anemia before planned surgery.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Grade 1C. "To avoid causing unnecessary delay to patients, anaemia screening should take place when referral for surgery is first made, in order to allow investigation and correction if appropriate".

Strong (Grade 1): There is confidence in the balance of risk or burden versus benefit. Grade 1 recommendations may be applied to most patients.

(C) Further research is likely to have an important impact on confidence in the effect estimate and is likely to change the estimate. Current evidence from observational studies, case series or expert opinion.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Weak (Grade 2): The balance between benefit and burden of therapy is less clear. Grade 2 recommendations require judicious application to individual patients.

(A) Further research is unlikely to change confidence in the estimate of effect. Current evidence derived from randomized clinical trials without important limitations.

(B) Further research may impact on confidence in the estimate of effect and may change the estimate. Current evidence derived from randomized clinical trials with important limitations (e.g. inconsistent results, imprecision or potential bias), or very strong evidence from observational studies (e.g. consistent estimates of the magnitude of a treatment effect or demonstration of a dose-response gradient).

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 2009 – September 2014

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*) 1 systematic review, 2 large studies (type unstated), 42 observational studies.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Grade is C, which is low-quality evidence from RCTs with major limitations or observational studies.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Grade is 1, indicating that benefit exceeds any risk

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Unstated

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

None

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, and other relevant sources including professional association websites and the National Guideline Clearinghouse was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

Society for the Advancement of Blood Management: “Patients who are having a procedure for which preoperative screening is required are identified at least three to four weeks prior to surgery to allow sufficient time to diagnose and manage anemia, unless the surgery is of an urgent nature and must be performed sooner.” (Standard 6.2) SABM Administrative and Clinical Standards for Patient Blood Management Programs, Third Edition. Unpublished work, 2014. Downloaded from www.SABM.org on April 9, 2016.

American Red Cross: “Preoperative assessment and efforts to reduce the RBC transfusion requirement in the perioperative period include the evaluation and treatment of anemia prior to surgery and the evaluation for discontinuation or replacement of anticoagulant and antiplatelet medications …for a sufficient time prior to surgery in consultation with the prescribing physician.” A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.

New York State Department of Health: “Careful evaluation of pre-existing anemia and its treatment prior to surgery are an effective strategy for reducing surgical transfusion requirements.” New York State Council on

Human Blood and Transfusion Services. Guidelines for Transfusion Options and Alternatives, 2010. Downloaded from www.wadswoth.org/labcert/blood_tissue July 2015.

"Recommendation 1. Whenever clinically feasible, elective surgical patients should have a hemoglobin level tested a minimum of 30 days before the scheduled surgical procedure." Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. *Anesth Analg* 2005;101:1858-61.

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_01_evidence_attachment-635993510467421034.docx

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Researchers have shown that preoperative hemoglobin (hgb) and hematocrit can be used as predictors of outcome for specific types of patients such as cardiac artery bypass graft or orthopedic surgery. 1 Preoperative anemia and exposure to allogeneic transfusion are associated with increased morbidity and mortality after surgery4. Previously undiagnosed anemia was identified in 5% - 75% of elective surgery patients in certain populations and a national audit demonstrated that 35% of patients scheduled for joint replacement therapy have a hgb. < 13 g/dL on preadmission testing.2 In the elderly (>65 yr. old), the prevalence of anemia as defined by the World Health Organization (WHO) is 11% and 10.2% for men and women, respectively.3 Blood transfusions are associated with several postsurgical complications, including surgical site infections, pneumonia, slower wound healing, prolonged ventilator use and increased length of stay.4

Development of formal protocols for preoperative testing of Hgb for potentially high blood loss elective surgeries could be used to identify and intervene for optimal management of blood resources. Early recognition of anemia offers patients an opportunity to receive the most appropriate transfusion-sparing strategy and avoid the risk of a potential transfusion. A panel of multidisciplinary physicians developed a clinical care pathway for anemia management in the elective surgical patients for whom blood transfusion is a probability (defined as any procedure for which a preoperative blood type and crossmatch is requested). They recommend that "Whenever clinically feasible, elective surgical patients should have a Hgb level tested a minimum of 30 days before the scheduled surgical procedure".2 Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need.4

1. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
2. Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005;101:1858-61.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. Br. Journ. Anesthesia, 106 (1): 13-22 (2011).

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included).*

This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.
No performance data are yet available.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
There are no studies in the literature indicating when pre-operative anemia screening is performed. In a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 118 of the 141 respondents (81%) indicated that there was a gap in practice; 6 were not sure, and 17 reported no gap. Of the 118, most indicated that pre-operative anemia screening was done 3 or 4 days in advance of the elective surgical procedure. (Unpublished data.)

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
No disparity data are available.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparity data are available in the literature.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
A. Incidence of preoperative anemia –
I. Incidence of anemia increases with age but varies by subpopulation.
a) Community-dwelling, >65 years old - <10%
b) Frail nursing home resident - >48%
c) Surgical population – 5% to 75%
d) Octogenarian, elective cardiac surgery – 49.4%1
e. 7% of 9,462 patients undergoing total hip or total knee replacement2
f. >65 years old – 11% women, 10.2% men (NHANES Study)3
g. Elective orthopedic surgery – 35%4
B. Preoperative Anemia as a risk factor
I. For blood transfusion after surgery –
a) 13 references (12 articles, one literature review) document increased rate of perioperative blood transfusion when preoperative anemia is present.5
b) 1 study of 296 elective orthopedic surgeries indicated that preoperative hemoglobin levels and patient weight were shown to predict the need for blood replacement after hip and knee replacement.7
c) 1 systematic literature review of 29 included citations demonstrated that low hemoglobin and patient age were consistent risk factors for blood transfusion in orthopedic surgery8
d) In a cohort study of 239 patients scheduled for transcatheter aortic valve implantation (TAVI), 62.3% were found to be anemic pre-procedurally and were referred to a blood conservation clinic (BCC) where they received a regimen of IV iron, oral iron, or epoetin alfa. Rates of transfusion in this cohort of 60 patients were assessed and compared with transfusion rates for TAVI patients prior to the initiation of the program. Implementation of the BCC was associated with a substantial decrease in the average blood transfusion

rate from 33.3% before program initiation to 15.3% after implementation (P < 0.001). After adjusting for baseline hemoglobin values and comorbidities, being assessed at the BCC was strongly associated with a reduction in the need for transfusion (odds ratio, 0.28; 95% confidence interval, 0.11-0.69; P ¼ 0.006).10

e) 9,482 patients undergoing hip or knee arthroplasty were evaluated in a multicenter study conducted in 1996-1997. Patients who had a baseline anemia (<13.0 G) had a higher prevalence of transfusions than did those who did not have anemia.11

f) A placebo-controlled, double-blind trial enrolling 316 patients scheduled for major, elective orthopedic hip or knee surgery who were expected to require 2 2 units of blood and who were not able or willing to participate in an autologous blood donation program examined the efficacy of Epogen treatment in reducing use of perioperative blood transfusion. Based on previous studies which demonstrated that pretreatment hemoglobin is a predictor of risk of receiving transfusion, patients were stratified into one of three groups based on their pretreatment hemoglobin [-< 10 (n = 2) > 10 to 5 13 (n = 96), and > 13 to I 15 g/dL (n = 218)] and then randomly assigned to receive 300 Units/kg EPOGENQ 100 Units/kg EPOGEN@ or placebo by SC injection for 10 days before surgery, on the day of surgery, and for 4 days after surgery. All patients received oral iron and a low-dose post-operative warfarin. Treatment with EPOGENB 300 Units/kg significantly (p = 0.024) reduced the risk of allogeneic transfusion in patients with a pretreatment hemoglobin of > 10 to _< 13 g/dL; 5/31 (16%) of EPOGENB 300 Units/kg, 6126 (23%) of EPOGEN@ 100 Units/kg, and 13/29 (45%) of placebo treated patients were transfused. There was no significant difference in the number of patients transfused between EPOGENB (9% 300 Units/kg, 6% 100 Units/kg) and placebo (13%) in the > 13 to I 15 g/dL hemoglobin stratum. There were too few patients in the I 10 g/dL group to determine if EPOGEN@ is useful in this hemoglobin strata. In the > 10 to I 13 g/dL pretreatment stratum, the mean number of units transfused per EPOGENQ-treated patient (0.45 units blood for 300 Units/kg, 0.42 units blood for 100 Units/kg) was less than the mean transfused per placebo-treated patient (1.14 units) (overall p = 0.028). In addition, mean hemoglobin, hematocrit and reticulocyte counts increased significantly during the pre-surgery period in patients treated with EPOGEN.12

II. For increased perioperative mortality

a) 1 study of 1958 Jehovah's Witness surgical patients – Hgb of <10 gm associated with significant increase in perioperative mortality6

IV. For all adverse outcomes

a) 1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that

1. The prevalence of preoperative anemia was 21-56%

2. Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.9

C. Transfusion as a high volume procedure:

Agency for Healthcare Research and Quality (AHRQ): Blood transfusion was the most common of all listed procedures performed during hospitalizations in 2010 (11 percent of stays with a procedure); the rate of hospitalization with blood transfusion has more than doubled since 1997. The percentage change in rate of all stays with a blood transfusion from 1997 – 2010 is 126%.13.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).

2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.

3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).

4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61

5. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". Ann Thorac Surg 2007;83: 527 – 86.

6. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

7. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. The Journal of Bone and Joint Surgery. Volume 84-A – Number2 – February 2002.

8. Barr PJ et al. Drivers of Transfusion Decision Making and Quality of the Evidence in Orthopedic Surgery: A Systematic Review of the Literature. Transfusion Medicine Reviews, Vol 25 No. 4 (October), 2011 pp. 304 – 316

9. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010

10. Shuvy M, et al. Preprocedure Anemia Management Decreases Transfusion Rates in Patients Undergoing Transcatheter Aortic Valve Implantation. Canadian Journal of Cardiology (2016) Article in press.

11. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.

12. deAndrade JH, Jove M. Baseline Hemoglobin as a Predictor of Risk of Transfusion and Response to Epoetin alfa in Orthopedic Surgical Patients. Am J of Orthoped. 1996;25(8): 533-542.
13. Most Frequent Procedures Performed in U.S. Hospitals, 2010. Healthcare Cost and Utilization project, statistical brief, February 2013, AHRQ

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not a PRO-PM.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

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

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-01_PreopAnemiaScreen.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PreopAnemiaScreen_v4_3_Thu_May_26_11.06.21_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients with preoperative anemia screening done in the window between 45 and 14 days prior to the surgery start date.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target*

process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Hemoglobin and hematocrit level drawn is represented as a code from the following value set and associated QDM datatype:
* "Laboratory Test, Performed: Hemoglobin Blood Serum Plasma" using "Hemoglobin Blood Serum Plasma LOINC Value Set (2.16.840.1.113762.1.4.1104.4)
Date of the elective surgical procedure is represented by a code from the following value set and associated QDM datatype:
* "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients age 18 and older with a length of stay less than or equal to 120 days who undergo selected elective surgical procedures

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
* "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
Selected elective surgical procedures are represented by a code from the following value set and associated QDM datatype:
* "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
• Patients whose surgical procedure is performed to address a traumatic injury • * Patients with a solid organ transplant recorded <=48 hours prior to the encounter or during the encounter

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Traumatic injury is represented by a code from the following value set and associated QDM datatype:
* Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
Solid organ transplant is represented by a code from the following value set and associated QDM datatype:

* "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
This measure is not stratified.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not a PRO-PM, measure is not based on a survey.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
2a. Reliability – See attached Measure Testing Submission Form **2b. Validity – See attached Measure Testing Submission Form** PBM01_CMS503v0_Bonnie_Export.xlsx,PBM_01_testing_form_for_trial_use-635999538879850795.docx

National Quality Forum

Measure Testing Form for Trial Approval Program

Measure Title: PBM-01: Preoperative Anemia Screening
Date of Submission: 5/31/2016
Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submission that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide preliminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the measure meets reliability and validity must be in this form

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*including questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

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

It is important that the measure developer use a data set to conduct preliminary testing in order to evaluate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

The Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.

Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 43T	☐ other:

1.4. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

22 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions.

All 22 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid. For further information on the characteristics of the patients included in the analysis, please refer to the attached BONNIE testing spreadsheet.

If the Bonnie testing tool was used to provide a sample data set, please refer to the guidance for Bonnie testing found at this link: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80307 Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. Refer to this link for an example of formatting Bonnie results: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=81576

Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.

Please refer to the attached BONNIE testing spreadsheet.

RELIABILITY AND VALIDITY ASSESSMENTS

Note: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) **Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program.**

Include descriptions of:

Inter-abstractor reliability, and data element reliability of all critical data elements

Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. BONNIE testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-01.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING

Numerator clearly describes the activity being measured	4.48
Denominator clearly describes the activity being measured	4.48
Numerator inclusions clear and appropriate	4.48
Denominator inclusions clear and appropriate	4.48
Numerator exclusions clear and appropriate	4.48
Denominator exclusions clear and appropriate	4.48
Accurately assesses the process of care to which it is addressed	4.28

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Initial Population and Denominator test cases positively test to ensure that only patients >= 18 years of age who have a surgical procedure performed <=48 hours prior to the inpatient encounter or during the inpatient encounter are included. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patients who have a surgical procedure at 49 hours and 48 hours prior to the start of the encounter. Patients who have a surgical procedure 48 hours prior to the start of the encounter were included in the denominator, while patients with a surgical procedure at 49 hours prior to the encounter were not.

Numerator test cases positively test to ensure patients who have a hemoglobin result recorded <= 45 days and >=14 days prior to the start of surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in which hemoglobin results were recorded >45 days prior to surgery or after surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures or encounter diagnoses. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-01 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance**

results? (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid organ transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PBM01_NQF_Measure_Feasibility_Assessment_Report-635999565863927224.docx

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
n/a

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.
Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)
There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Planned	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose

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

This is a new measure.

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

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences identified during testing.

5. Comparison to Related or Competing Measures

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

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

No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
There are no competing measures.

Appendix

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

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content

oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List

Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health
Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care
Medicine, Division of Vascular, Thoracic, Transplant Anesthesia

Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.
Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center

Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America
Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic
Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic
Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center
Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:** Measure specifications are in the Public Domain
LOINC(R) is a registered trademark of the Regenstrief Institute.
This material contains SNOMED Clinical Terms (R) (SNOMED CT(c)) copyright 2004-2014 International Health Terminology Standards Development Organization. All rights reserved.
Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-01: Preoperative Anemia Screening

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility

throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements which test sites deemed to be difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"
3. "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"
4. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
5. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Initial Population and Denominator Data Elements

Data elements 1- “"Encounter, Performed: Encounter Inpatient" and 3- "Procedure, Performed: Selected Elective Surgical Procedures" are used to define the initial the initial population and denominator of this measure. On the feasibility scorecard, hospitals rated these data elements as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year time frame outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Four out of five hospitals rated capture of data element 3 as feasible or highly feasible, represented as a score of 2 or 3 out of 3. Facilities rating the data element as a 2 cited variation in clinical workflow and adoption of new technology as reasons for the lower rating. One site rated current state feasibility as a 1, as it did not currently have an interface between the OR scheduling system where this information was captured and the certified EHR technology. This site had plans to transition to an interfaced OR module in 1-2 years. All sites rated the future state as highly feasible.

Numerator Data Element

Data element 2- "Laboratory Test, Performed: Hemoglobin blood serum plasma" is used to define the numerator population for this measure. Specifically, the numerator evaluates whether patients had a laboratory test performed assessing hemoglobin level in the window between 45 and 14 days before the start of the elective surgical procedure.

Hospitals reported a feasibility score of 2 for workflow and data availability for this data element. While lab results are routinely captured as structured data, limited interoperability between hospitals and their community partners, such as clinics and lab centers, limits the availability of structured data for lab results occurring prior to the hospital encounter. Hospitals noted that many external results are received via fax, or as an electronic document, rather than in a format that can be structured and encoded in the EHR.

The Joint Commission views the Approval for Trial Use designation as an opportunity to work with developers and implementers to further explore methods to improve feasibility for this data element.

Please refer to Appendix A for further findings related to this data element.

Denominator Exclusions

Patients with documentation of data elements 4- “"Procedure, Performed: Solid Organ Transplant" and 5- "Diagnosis: Traumatic Injury" are excluded from the measure.

Feasibility for data element 4- "Procedure, Performed: Solid Organ Transplant" was found to be comparable to data element 3- "Procedure, Performed: Selected Elective Surgical Procedures.” These data elements are both found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform organ transplant, which would not use this data element.

All hospitals rated data element 8- "Diagnosis: Traumatic Injury," as highly feasible. Discussion around this data element suggested that while missing data may occur due to physician practice related to updating the problem list, the functionality to support collection of this data element was well established.

Conclusion

Hospitals completing the feasibility scorecard largely reported the data elements required to calculate this measure to be feasible or highly feasible in the current state, with the exception of the numerator data element representing hemoglobin results. Capture of hemoglobin results will require improvements in interoperability or workarounds to support data collection. Approval for Trial Use status will support The Joint Commission’s efforts to further test this measure.

Appendix A: Feasibility Scorecard Findings for External Hemoglobin Results

Site		Workflow		Data Availability		Data Accuracy		Data Element Definition		Data Standard	
		Score	Comments	Score	Comments	Score	Comments	Score	Comments	Score	Comments
1	Current	3		3		3		3		3	
	Future (3-5 years)	3		3		3		3		3	
2	Current	2		2	Sometimes procedures are performed without a hgb, and not always captured 14-45 days in advance	2		**3**		**3**	
	Future (3-5 years)	3	.	3		3		**3**		**3**	
3	Current	2	Not currently captured within time frame specified.	1	Never captured in time frame, most are scanned documents as they are not performed at facility	3		3		3	
	Future (3-5 years)	2	External results may or may not interface	2		3		3		3	
4	Current	2	Some results are scanned in from outside systems	2	Some results are scanned	3	Captured data is accurate	3		3	Not reported for scanned results
	Future (3-5 years)	3	Have educational plans in place to improve practice	2	Expect care model and technology will improve data availability, but still expect data will not always be available	3		3		3	
5	Current	2		2		3		3		3	
	Future (3-5 years)	2		2		3		3		3	

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3021
De.2. **Measure Title:** PBM-05: Blood Usage, Selected Elective Surgical Patients
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure assesses the proportion of selected elective surgical patients age 18 and over who had a timely preoperative anemia screening and subsequent perioperative transfusion. Since preoperative anemia is a predictor of perioperative transfusion, this measure can identify records of patients needing further review for uncorrected preoperative anemia or other blood management measures, such as a restrictive transfusion strategy or cell salvage, that should have been taken to avoid transfusion.
1b.1. **Developer Rationale:** Research shows that correction of preoperative anemia significantly reduces perioperative transfusion.1,2,3 Preoperative anemia is also associated with adverse outcomes after surgery.4,5 The rationale for the measure is identification of patients who had timely preoperative anemia screening but underwent elective surgery with uncorrected anemia, causing a perioperative transfusion to be administered. The measure can also identify opportunities for other methods of blood conservation, such as cell salvage and restrictive transfusion strategy that may have been missed. Over time, the proportion of patients in the numerator should decrease, leading to better patient outcomes and conservation of blood resources.

1. American Red Cross. A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.
2. Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005;101:1858-61, p. 1860
3. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". Ann Thorac Surg 2007;83: 527 – 86.
4. Fowler AJ et al. Meta-analysis of the association between preoperative anaemia and mortality after surgery. Br J Surg 2015 Oct;102(11):1314-24.
5. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

S.4. **Numerator Statement:** Patients who had a non-autologous whole blood or non-autologous packed red blood cell transfusion administered in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.
S.7. **Denominator Statement:** Selected elective surgical patients age 18 and older who had a preoperative anemia screening in the time window between 45 and 14 days before surgery start date.
S.10. **Denominator Exclusions:**

- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients with sickle cell disease or hereditary hemoglobinopathy
- Patients who refuse blood transfusion.
- Patients who receive an autologous blood transfusion

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

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

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

The developer provides the following path to support the relationship between transfusion and outcomes:

1. Processes: Optimized preoperative hemoglobin level and restrictive transfusion strategy
2. Elective surgical procedure performed
3. Reduced rate of blood transfusion
4. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

The rationale for the measure is supported by five clinical guideline recommendations:

1. Network for Advancement of Transfusion Alternatives*: We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl−1, male ≥13 g dl−1), according to the WHO criteria. This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anemia and associated clinical conditions will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.* (Grade 2C – Weak recommendation/low quality evidence). A summary of the QQC is partially provided.
2. AABB*:*
   *a): The AABB recommends adhering to a restrictive transfusion strategy (7 to 8 g/dL) in hospitalized, stable patients. (*Grade: strong recommendation; high-quality evidence)
   b): The AABB suggests adhering to a restrictive strategy in hospitalized patients with preexisting cardiovascular disease and considering transfusion for patients with symptoms or a hemoglobin level of 8 g/dL or less. (Grade: weak recommendation; moderate-quality evidence)
   c): The AABB cannot recommend for or against a liberal or restrictive transfusion threshold for hospitalized, hemodynamically stable patients with the acute coronary syndrome. (uncertain recommendation; very low-quality evidence)
   A summary of the QQC is not provided for these recommendations

3. Society of Thoracic Surgeons/ The Society of Cardiovascular Anesthesiologists: *With hemoglobin levels below 6 g/dL, red blood cell transfusion is reasonable since this can be lifesaving. Transfusion is reasonable in most postoperative patients whose hemoglobin is less than 7 g/dL but no high level evidence supports this recommendation.* (Level of evidence C) A summary of the QQC is not provided.
4. American Red Cross: *A restrictive RBC transfusion strategy (Hgb 7–8 g/dL trigger) is recommended in stable hospitalized patients.* (No grade assignment). A summary of the QQC is not provided.
5. Society of Critical Care Medicine: *A "restrictive" strategy of RBC transfusion (transfusion when Hb <7 g/dL) is as effective as a "liberal" strategy (transfusion when Hb < 10 g/dL) in critically ill patients with hemodynamically stable anemia, except possibly in patients with acute myocardial ischemia. (*Level 1: The recommendation is convincingly justifiable based on the available scientific information alone) A summary of the QQC is not provided.

The developer provided additional citations as sources of evidence for this measure. The most relevant piece of literature provided for this measure is reference #13, summarized below:

- *1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that*
    a. *The prevalence of preoperative anemia was 21-56%*
    b. *Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.*
- Reference #13 is a systematic review but not graded, and a summary of the Quantity, Quality and Consistency of the body of evidence were not provided.

Exception to evidence

N/A

Guidance from the Evidence Algorithm

Process measure based on SR/grading (Box 3) → QQC of the body of evidence from relevant SR provided was not provided (Box 4) → No grading of evidence and summary of QQC not provided (Box 6) → LOW

Note: If a summary of the QQC was provided, the highest possible rating would be: MODERATE.

Questions for the Committee:

For process measures:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

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

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

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

Although there is no performance data on the measure as specified, the developer listed data from the literature that indicates opportunity for improvement on the prevalence of anemia.

Disparities

- The developed indicated that no disparity data are available.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Does a gap in care exist for transfusion rate in elective surgery?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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

Committee pre-evaluation comments

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

1a. Evidence to Support Measure Focus

- Complex measure, but getting at an actual important outcome. Could be combined with the first measure as one assessment.
  Might want to rephrase as a positive measure.
  Same issue with surgeries selected.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients who had a non-autologous whole blood or non-autologous packed red blood cell transfusion administered in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.
- Denominator Statement: Selected elective surgical patients age 18 and older who had a preoperative anemia screening in the time window between 45 and 14 days before surgery start date.
- Denominator Exclusions:
  - Patients whose surgical procedure is performed to address a traumatic injury
  - Patients under 18
  - Patients who have a solid organ transplant
  - Patients with sickle cell disease or hereditary hemoglobinopathy
  - Patients who refuse blood transfusion.
  - Patients who receive an autologous blood transfusion
- Level of Analysis: Facility
- Care Setting: Hospital/Acute Care Facility
- No risk adjustment or risk stratification

eMeasure Technical Advisor(s) review

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM

Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously; Bonnie results included with submission
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 24 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications are consistent with evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.35
Denominator clearly describes the activity being measured	4.36

Numerator inclusions clear and appropriate	4.31
Denominator inclusions clear and appropriate	4.23
Numerator exclusions clear and appropriate	4.34
Denominator exclusions clear and appropriate	4.33
Accurately assesses the process of care to which it is addressed	4.00

This measure is being considered for trial use, thus full validity testing results are not expected and the Committee will not vote on this criterion.

2b3-2b7. Threats to Validity

2b3. Exclusions:

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid organ transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury or sickle cell disease
- Patients who refuse transfusion

- ***Questions for the Committee:*** *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

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

Committee pre-evaluation comments

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

Criterion 3. Feasibility

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

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

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

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

Committee pre-evaluation comments

Criteria 3: Feasibility

Criterion 4: Usability and Use

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms N/A

Feedback :

None identified

Questions for the Committee:

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ High ☒ Moderate ☐ Low ☐ Insufficient
Committee pre-evaluation comments **Criteria 4: Usability and Use**

Criterion 5: Related and Competing Measures
Related or competing measures N/A **Harmonization** N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 42T

Measure Title: PBM-05: Blood Usage in Selected Elective Surgical Patients

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 42T

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus

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

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

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

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

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

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

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

Outcome

☐ Health outcome: 42T

☐Patient-reported outcome (PRO): 42T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 42T

☒ Process: Optimized preoperative hemoglobin level and restrictive transfusion strategy.

☐ Structure: 42T

☐ Other: 42T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

5. Processes: Optimized preoperative hemoglobin level and restrictive transfusion strategy
6. Elective surgical procedure performed
7. Reduced rate of blood transfusion
8. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

Guideline #1.

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. *Br. Journ. Anesthesia*, 106 (1): 13-22 (2011).

http://bja.oxfordjournals.org/content/106/1/13.full

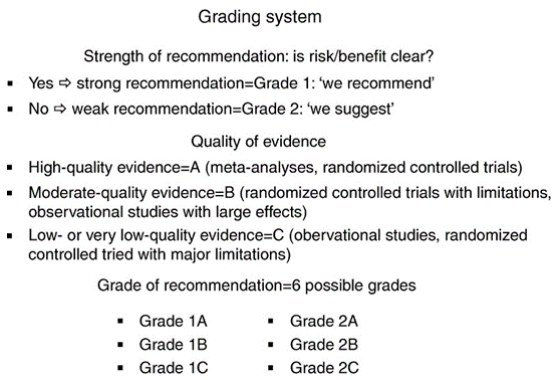
1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 2: We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl^{-1}, male ≥13 g dl^{-1}), according to the WHO criteria (Grade 2C).

This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anaemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade 2C



1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

See 1a.4.3

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as 1a.4.1

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Guideline #1.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Detection, evaluation, and management of preoperative anemia in elective orthopedic surgery.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Grade C – low-quality evidence (observational studies, randomized control trials with major limitations).

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.4.3

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1966 – January 2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

5 observational studies, 3 cohort studies, 1 meta-analysis, 1 systematic literature review.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not stated in citation; in review of studies appears that 3 are small cohort studies.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Unstated in citation

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

None.

Guideline #2.

<u>AABB</u>

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Carson JL, Grossman BJ, Kleinman S, Tinmouth at, et al. Red Blood Cell Transfusion: A Clinical Practice Guideline From the AABB. *Ann Intern Med.* 2012;157(1):49-58.

http://annals.org/article.aspx?articleid=1206681

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 1: The AABB recommends adhering to a restrictive transfusion strategy (7 to 8 g/dL) in hospitalized, stable patients.

Recommendation 2: The AABB suggests adhering to a restrictive strategy in hospitalized patients with preexisting cardiovascular disease and considering transfusion for patients with symptoms or a hemoglobin level of 8 g/dL or less.

Recommendation 3: The AABB cannot recommend for or against a liberal or restrictive transfusion threshold for hospitalized, hemodynamically stable patients with the acute coronary syndrome.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Recommendation 1: Grade: strong recommendation; high-quality evidence.

Recommendation 2: Grade: weak recommendation; moderate-quality evidence.

Recommendation 3: Grade: uncertain recommendation; very low-quality evidence.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

The strength of recommendations (for or against intervention) is graded as "strong" (indicating judgment that most well-informed people will make the same choice; "We recommend . . . "), "weak" (indicating judgment that a majority of well-informed people will make the same choice, but a substantial minority will not; "We suggest . . . "), or "uncertain" (indicating that the panel made no specific recommendation for or against interventions; "We cannot recommend . . . ").

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Clinical Practice Guideline, AABB:

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The GRADE system (39) uses the following 4 ratings for quality of evidence:

"High" indicates considerable confidence in the estimate of effect. The true effect probably lies close to the estimated effect, and future research is unlikely to change the estimate of the health intervention's effect.

"Moderate" indicates confidence that the estimate is close to the truth. Further research is likely to have an important effect on confidence in the estimate and may change the estimate of the

health intervention's effect.

"Low" indicates that confidence in the effect is limited. The true effect may differ substantially from the estimate, and further research is likely to have an important effect on confidence in the estimate of the effect and is likely to change the estimate.

"Very low" indicates little confidence in the effect estimate. Any estimate of effect is very uncertain.

Guideline #3

Society of Thoracic Surgeons/ The Society of Cardiovascular Anesthesiologists

1a.4.1. Guideline citation (*including date*) and URL for guideline (*if available online*):

Ferraris V, Brown JR, Despotis GJ, Hammon JW, et al. 2011 Update to the Society of Thoracic Surgeons and the Society of Cardiovascular Anesthesiologists Blood Conservation Clinical Practice Guidelines.

Ann Thorac Surg 2011;91:944-82.

http://www.annalsthoracicsurgery.org/article/S0003-4975(10)02888-2/pdf

1a.4.2. Identify guideline recommendation number and/or page number and quote verbatim, the specific guideline recommendation.

(No Number Table 2): "With hemoglobin levels below 6 g/dL, red blood cell transfusion is reasonable since this can be lifesaving. Transfusion is reasonable in most postoperative patients whose hemoglobin is less than 7 g/dL but no high level evidence supports this recommendation. (Level of evidence C)."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Class IIA **CLASS IIa,** *Benefit >> Risk*
Additional studies with focused objectives needed. **IT IS REASONABLE** to perform procedure/administer treatment.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

CLASS I, *Benefit >>> Risk*
Procedure/Treatment **SHOULD** be performed/administered

CLASS IIb, *Benefit ≥ Risk*
Additional studies with broad objectives needed; additional registry data would be helpful.
Procedure/Treatment **MAY BE CONSIDERED**

CLASS III, *Risk ≥ Benefit*
Procedure/Treatment should **NOT** be performed/administered **SINCE IT IS NOT HELPFUL AND MAY BE HARMFUL**

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*):

Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines

http://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/documents/downloadable/ucm_319826.pdf

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → *complete section 1a.7*

☒ No → *report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7*

Guideline #4

American Red Cross Guideline

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Vassallo R, et al. A Compendium of Transfusion Practice Guidelines, Second Edition, 2013. American Red Cross, page 8.
http://www.redcrossblood.org/sites/arc/files/59802_compendium_brochure_v_6_10_9_13.pdf

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page 15: A restrictive RBC transfusion strategy (Hgb 7–8 g/dL trigger) is recommended in stable hospitalized patients.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

No grade assignment

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

n/a

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

n/a

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

Guideline #5

Society of Critical Care Medicine:

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Napolitano L, et al. Clinical Practice Guideline: Red blood cell transfusion in adult trauma and critical care. Crit Care Med 2009 Vol 37, No 12.

http://journals.lww.com/ccmjournal/Abstract/2009/12000/Clinical_practice_guideline__Red_blood_cell.19.aspx

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page 3127, recommendation 3: "A "restrictive" strategy of RBC transfusion (transfusion when Hb <7 g/dL) is as effective as a "liberal" strategy (transfusion when Hb < 10 g/dL) in critically ill patients with hemodynamically stable anemia, except possibly in patients with acute myocardial ischemia".

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Level 1. The recommendation is convincingly justifiable based on the available scientific information alone. This recommendation is usually based on Class I data, however strong Class II evidence may form the basis for a Class 1 recommendation.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Level 2. The recommendation is reasonably justifiable by available scientific evidence and strongly supported by expert opinion. This recommendation is usually supported by Class II data or a preponderance of Class III evidence.

Level 3. The recommendation is supported by available data but adequate scientific evidence is lacking.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same, p. 3126.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, and other relevant sources including professional association websites, The Cochrane Library, the National Guideline Clearinghouse, and other sources was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

1. American Red Cross: "Preoperative assessment and efforts to reduce the RBC transfusion requirement in the perioperative period include the evaluation and treatment of anemia prior to surgery and the evaluation for discontinuation or replacement of anticoagulant and antiplatelet medications …for a sufficient time prior to surgery in consultation with the prescribing physician." A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.

2. Society for Blood Management: The panel further recommended that the patient's target hemoglobin before elective surgery should be within the normal range (normal female >12 g/dL, normal male >13 g/dL). Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. *Anesth Analg* 2005;101:1858-61, p. 1860.

3. "Preoperative anaemia is associated with poor outcomes after surgery" Fowler AJ et al. Meta-analysis of the association between preoperative anaemia and mortality after surgery. *Br J Surg* 2015 Oct;102(11):1314-24.

METHODS:

A systematic review and meta-analysis of observational studies exploring associations between preoperative anaemia and postoperative outcomes was performed. Studies investigating trauma, burns, transplant, paediatric and obstetric populations were excluded. The primary outcome was 30-day or in-hospital mortality. Secondary outcomes were acute kidney injury, stroke and myocardial infarction. Predefined analyses were performed for the cardiac and non-cardiac surgery subgroups. A post hoc analysis was undertaken to evaluate the relationship between anaemia and infection. Data are presented as odds ratios (ORs) with 95 per cent c.i.

RESULTS:

From 8973 records, 24 eligible studies including 949 445 patients were identified. Some 371 594 patients (39·1 per cent) were anaemic. Anaemia was associated with increased mortality (OR 2·90, 2·30 to 3·68; I(2) = 97 per cent; P < 0·001), acute kidney injury (OR 3·75, 2·95 to 4·76; I(2) = 60 per cent; P < 0·001) and infection (OR 1·93, 1·17 to 3·18; I(2) = 99 per cent; P = 0·01). Among cardiac surgical patients, anaemia was associated with stroke (OR 1·28, 1·06 to 1·55; I(2) = 0 per cent; P = 0·009) but not myocardial infarction (OR 1·11, 0·68 to 1·82; I(2) = 13 per cent; P = 0·67). Anaemia was associated with an increased incidence of red cell transfusion (OR 5·04, 4·12 to 6·17; I(2) = 96 per cent; P < 0·001). Similar findings were observed in the cardiac and non-cardiac subgroups.

CONCLUSION:

Preoperative anaemia is associated with poor outcomes after surgery, although heterogeneity between studies was significant. It remains unclear whether anaemia is an independent risk factor for poor outcome or simply a marker of underlying chronic disease. However, red cell transfusion is much more frequent amongst anaemic patients.

4. British Committee for Standards in Haemotology: Recommendation: "Healthcare pathways should be structured to ensure anaemia screening and correction before surgery." Kotze A, Harris A, Baker C, Iqbal T, eta l. British Committee for Standards in Haemotology Guidelines on the Identification and Management of Pre-Operative Anemia. *British Journal of Haemotology* Volume 171, Issue 3: November 2015 pages 322-331.

5. Society for the Advancement of Blood Management: "Patients who are having a procedure for which preoperative screening is required are identified at least three to four weeks prior to surgery to allow sufficient time to diagnose and manage anemia, unless the surgery is of an urgent nature and must be performed sooner." (Standard 6.2) SABM Administrative and Clinical Standards for Patient Blood Management Programs, Third Edition. Unpublished work, 2014. Downloaded from www.SABM.org on April 9, 2016.

6. New York State Department of Health: "Careful evaluation of pre-existing anemia and its treatment prior to surgery are an effective strategy for reducing surgical transfusion requirements." New York State Council on Human Blood and Transfusion Services. Guidelines for Transfusion Options and Alternatives, 2010. Downloaded from www.wadswoth.org/labcert/blood_tissue July 2015.

7. 13 references (12 articles, one literature review) document increased rate of perioperative blood transfusion when preoperative anemia is present. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". *Ann Thorac Surg* 2007;83: 527 – 86

8. 1 study of 296 elective orthopedic surgeries indicated through multivariate analysis that a significant relationship existed only between the need for transfusion and the preoperative hemoglobin level (p+ 0.00001) after hip and knee replacement. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. *The Journal of Bone and Joint Surgery*. Volume 84-A – Number2 – February 2002.

9. 1 systematic literature review of 29 included citations demonstrated that low hemoglobin and patient age were consistent risk factors for blood transfusion in orthopedic surgery. · Barr PJ et al. Drivers of Transfusion Decision Making and Quality of the Evidence in Orthopedic Surgery: A Systematic Review of the Literature. *Transfusion Medicine Reviews*, Vol 25 No. 4 (October), 2011 pp. 304 – 316.

10. In a cohort study of 239 patients scheduled for transcatheter aortic valve implantation (TAVI), 62.3% were found to be anemic pre-procedurally and were referred to a blood conservation clinic (BCC) where they received a regimen of IV iron, oral iron, or epoetin alfa. Rates of transfusion in this cohort of 60 patients were assessed and compared with transfusion rates for TAVI patients prior to the initiation of the program. Implementation of the BCC was associated with a substantial decrease in the average blood transfusion rate from 33.3% before program initiation to 15.3% after implementation (P < 0.001). After adjusting for baseline hemoglobin values and comorbidities, being assessed at the BCC was strongly associated with a reduction in the need for transfusion (odds ratio, 0.28; 95% confidence interval, 0.11-0.69; P ¼ 0.006. Shuvy M, et al. Preprocedure Anemia Management Decreases Transfusion Rates in Patients Undergoing Transcatheter Aortic Valve Implantation. *Canadian Journal of Cardiology* (2016) Article in press.

11. A placebo-controlled, double-blind trial enrolling 316 patients scheduled for major, elective orthopedic hip or knee surgery who were expected to require 2.2 units of blood and who were not able or willing to participate in an autologous blood donation program examined the efficacy of Epogen treatment in reducing use of perioperative blood transfusion. Based on previous studies which demonstrated that pretreatment hemoglobin is a predictor of risk of receiving transfusion, patients were stratified into one of three groups based on their pretreatment hemoglobin [-< 10 (n = 2) > 10 to 5 13 (n = 96), and > 13 to I 15 g/dL (n = 218)] and then randomly assigned to receive 300 Units/kg EPOGENQ 100 Units/kg EPOGEN@ or placebo by SC injection for 10 days before surgery, on the day of surgery, and for 4 days after surgery. All patients received oral iron and a low-dose post-operative warfarin. Treatment with EPOGENB 300 Units/kg significantly (p = 0.024) reduced the risk of allogeneic transfusion in patients with a pretreatment hemoglobin of > 10 to _< 13 g/dL; 5/31 (16%) of EPOGENB 300 Units/kg, 6126 (23%) of EPOGEN@ 100 Units/kg, and 13/29 (45%) of placebo treated patients were transfused. There was no significant difference in the number of patients transfused between EPOGENB (9% 300 Units/kg, 6% 100 Units/kg) and placebo (13%) in the > ***13*** to I ***15*** g/dL hemoglobin stratum. There were too few patients in the I 10 g/dL group to determine if EPOGEN@ is useful in this hemoglobin strata. In the > 10 to I 13 g/dL pretreatment stratum, the mean number of units transfused per EPOGENQ-treated patient (0.45 units blood for 300 Units/kg, 0.42 units blood for 100 Units/kg) was less than the mean transfused per placebo-treated patient (1.14 units) (overall p = 0.028). In addition, mean hemoglobin, hematocrit and reticulocyte counts increased significantly during the pre-surgery period in patients treated with EPOGEN. deAndrade JH, Jove M. Baseline Hemoglobin as a Predictor of Risk of Transfusion and Response to Epoetin alfa in Orthopedic Surgical Patients. *Am J of Orthoped*. 1996;25(8): 533-542.

12. Among 569 patients who underwent colorectal cancer surgery between 1998 and 2003, 32 anemic patients who received iron supplementation for at least 2 weeks preoperatively (group A) and 84 anemic patients who did not (group B) were studied.
There were no significant differences between groups A and B in age, sex, surgical technique, tumor stage, and operating time. Their Hgb and Hct values were similar at first presentation, but significantly different immediately before surgery (both P < 0.0001). There were no significant differences in intraoperative blood loss between the groups, but significantly fewer patients in group A needed an intraoperative blood transfusion (9.4% vs 27.4%, P < 0.05). Okuyama M et al. Preoperative iron supplementation and intraoperative transfusion during colorectal cancer surgery. *Surg Today*. 2005;35(1):36-40.

13. 1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that
c. The prevalence of preoperative anemia was 21-56%
d. Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.
Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. *Anesthesiology*, v. 113 No 2 August 2010.

14. A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Composite postoperative morbidity at 30 days was also higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

15. A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of men and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_05_evidence_attachment.docx

1b. Performance Gap

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

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

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Research shows that correction of preoperative anemia significantly reduces perioperative transfusion.1,2,3 Preoperative anemia is also associated with adverse outcomes after surgery.4,5 The rationale for the measure is identification of patients who had timely preoperative anemia screening but underwent elective surgery with uncorrected anemia, causing a perioperative transfusion to be administered. The measure can also identify opportunities for other methods of blood conservation, such as cell salvage and restrictive transfusion strategy that may have been missed. Over time, the proportion of patients in the numerator should decrease, leading to better patient outcomes and conservation of blood resources.

1. American Red Cross. A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.
2. Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005;101:1858-61, p. 1860
3. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". Ann Thorac Surg 2007;83: 527 – 86.
4. Fowler AJ et al. Meta-analysis of the association between preoperative anaemia and mortality after surgery. Br J Surg 2015 Oct;102(11):1314-24.
5. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

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

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

Incidence of preoperative anemia –

a. Incidence of anemia increases with age but varies by subpopulation.

i. Community-dwelling, >65 years old - <10%

ii. Frail nursing home resident - >48%

iii. Surgical population – 5% to 75%

iv. Octogenarian, elective cardiac surgery – 49.4%1

v. 7% of 9,462 patients undergoing total hip or total knee replacement2

vi. >65 years old – 11% women, 10.2% men (NHANES Study)3

vii. Elective orthopedic surgery – 35%4

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).

2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.

3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).

4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

b. Further evidence that preoperative anemia is prevalent:

i. A systematic review and meta-analysis of observational studies exploring associations between preoperative anaemia and postoperative outcomes was performed. From 8973 records, 24 eligible studies including 949?445 patients were identified. Some 371?594 patients (39·1 per cent) were anaemic. Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005;101:1858-61, p. 1860.

ii. In a cohort study of 239 patients scheduled for transcatheter aortic valve implantation (TAVI), 62.3% were found to be anemic pre-procedurally. Shuvy M, et al. Preprocedure Anemia Management Decreases Transfusion Rates in Patients Undergoing Transcatheter Aortic Valve Implantation. Canadian Journal of Cardiology (2016) Article in press.

iii. 1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that

1. The prevalence of preoperative anemia was 21-56%

2. Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.

Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.

iv. A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

v. A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of men and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.

In addition, in a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 118 of the 141 respondents (81%) indicated that there was a gap in practice; 6 were not sure, and 17 reported no gap. Of the 118, most indicated that pre-operative anemia screening was done 3 or 4 days in advance of the elective surgical procedure. Given that 3-4 days is an insufficient period of time to correct any anemia, we suspect a high incidence of patients undergoing elective surgery with uncorrected anemia.

c. Opportunity for restrictive transfusion strategy:

Shander et al. Appropriateness of Allogeneic Red Blood Cell Transfusion: The International Consensus Conference on Transfusion Outcomes. Transfusion Medicine Reviews, Vol 25, No 3 (July), 2011: pp 232-246.e53.

An international multidisciplinary panel of 15 experts reviewed 494 published articles and used the RAND/UCLA Appropriateness Method to determine the appropriateness of allogeneic red blood cell (RBC) transfusion based on its expected impact on outcomes of stable nonbleeding patients in 450 typical inpatient medical, surgical, or trauma scenarios. Panelists rated allogeneic RBC transfusion as appropriate in 53 of the scenarios (11.8%), inappropriate in 267 (59.3%), and uncertain in 130 (28.9%). Red blood cell

transfusion was most often rated appropriate (81%) in scenarios featuring patients with hemoglobin (Hb) level 7.9 g/dL or less, associated comorbidities, and age older than 65 years. Red blood cell transfusion was rated inappropriate in all scenarios featuring patients with Hb level 10 g/dL or more and in 71.3% of scenarios featuring patients with Hb level 8 to 9.9 g/dL. Conversely, no scenario with patient´s Hb level of 8 g/dL or more was rated as appropriate. Nearly one third of all scenarios were rated uncertain, indicating the need for more research. The observation that allogeneic RBC transfusions were rated as either inappropriate or uncertain in most scenarios in this study supports a more judicious transfusion strategy.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
n/a

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparity data hav been identified in the literature.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Incidence of preoperative anemia –
d. Incidence of anemia increases with age but varies by subpopulation.
i. Community-dwelling, >65 years old - <10%
ii. Frail nursing home resident - >48%
iii. Surgical population – 5% to 75%
iv. Octogenarian, elective cardiac surgery – 49.4%1
v. 7% of 9,462 patients undergoing total hip or total knee replacement2
vi. >65 years old – 11% women, 10.2% men (NHANES Study)3
vii. Elective orthopedic surgery – 35%4
Blood transfusion is the most common procedure performed during hospitalization,5

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1 Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.
5. Most Frequent Procedures Performed in U.S. Hospitals, 2010, Healthcare Cost and Utilization Project (HCUP). February 2013. Agency for Healthcare Research and Quality.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not a PRO-PM.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Overuse, Safety

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

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-05_BloodUsageinSESP.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** BloodUsageinSESP_v4_3_Wed_May_25_08.49.06_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had a non-autologous whole blood or non-autologous packed red blood cell transfusion administered in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Non-autologous whole blood or non-autologous packed red blood cell transfusion is represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set

(2.16.840.1.113762.1.4.1029.24)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Selected elective surgical patients age 18 and older who had a preoperative anemia screening in the time window between 45 and 14 days before surgery start date.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Inpatients age 18 and over are represented by a code from the following Value Set and associated QDM Datatype:
"Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
Selected elective surgical patients are represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"
Preoperative anemia screening is represented by a code from the following Value Set and associated QDM datatype:
"Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients with sickle cell disease or hereditary hemoglobinopathy
- Patients who refuse blood transfusion.
- Patients who receive an autologous blood transfusion

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Traumatic injury is represented by a code from the following Value Set and associated QDM datatype:
Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
Solid organ transplant is represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
Sickle cell disease or hereditary hemoglobinopathy is represented by a code from the following Value Set and associated QDM datatype:
Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"
Patients who refuse transfusion are represented by a code from the following Value Set and associated QDM datatype:
Procedure, Order not done: Patient Refusal" using "Patient Refusal SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.93)"
Patients who receive autologous blood are represented by a code from the following Valu Set and associated QDM datatype:
"Substance, Order: Autologous Blood Product" using "Autologous Blood Product SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.36)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
This measure is not stratified.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific*

Acceptability)
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Measure is not based on a survey; not a PRO-PM.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM05_CMS607v0_Bonnie_Export.xlsx,PBM_05_testing_form_for_trial_use.docx

National Quality Forum

Measure Testing Form for Trial Approval Program

Measure Title: PBM-05: Blood Usage in Selected Elective Surgical Patients

Date of Submission: 5/31/2016

Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submission that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide preliminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the the measure meets the reliability and validity must be in this form..

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*including questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

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

It is important that the measure developer use a data set to conduct preliminary testing in order to evaluate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can use can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

The Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.

Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.,*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 42T	☐ other:

1.4. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

24 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions.

All 24 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid. For further information on the characteristics of the patients included in the analysis, please refer to the attached BONNIE testing spreadsheet.

1.5. Please refer to the guidance for Bonnie testing found at this link. Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. *Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.*

Please refer to the attached BONNIE testing spreadsheet.

RELIABILITY AND VALIDITY ASSESSMENTS

Note: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program. Include descriptions of:

Inter-abstractor reliability, and data element reliability of all critical data elements

Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-05.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below

presents the average rating for these parameters for PBM-05: Blood Usage in Selected Elective Surgical Patients:

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.35
Denominator clearly describes the activity being measured	4.36
Numerator inclusions clear and appropriate	4.31
Denominator inclusions clear and appropriate	4.23
Numerator exclusions clear and appropriate	4.34
Denominator exclusions clear and appropriate	4.33
Accurately assesses the process of care to which it is addressed	4.00

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Initial Population and Denominator test cases positively test to ensure that only patients >= 18 years of age who have a surgical procedure performed <=48 hours prior to the inpatient encounter or during the inpatient encounter and who have a hemoglobin test performed in the time period of <=45 days and >=14 days prior to surgery are included. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patients who have a surgical procedure at 49 hours and 48 hours prior to the start of the encounter. Patients who have a surgical procedure 48 hours prior to the start of the encounter were included in the denominator, while patients with a surgical procedure at 49 hours prior to the encounter were not.

Numerator test cases positively test to ensure patients who have a blood transfusion are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in blood group antibody screens were recorded >45 days prior to surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures, encounter diagnoses, or refusal of blood transfusion. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-05 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid organ transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury or sickle cell disease
- Patients who refuse transfusion

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment **Attachment:** PBM05_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

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

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
No modifications have been made.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Usability and Use Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.
Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)
There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

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

4a. Accountability and Transparency

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

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

Planned	Current Use (for current use provide URL)
Public Reporting	

Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

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

This is a new measure.

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

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

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

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

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

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
n/a

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

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

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

5a. Harmonization

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

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

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

5b. Competing Measures

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

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

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

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission

Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630---

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List

Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health
Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care Medicine, Division of Vascular, Thoracic, Transplant Anesthesia

Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.

Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center

Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America

Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic

Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic

Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center

Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-05: Blood Usage in Selected Elective Surgical Patients

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements that were deemed to be more difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"
3. "Procedure, Order: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
4. "Procedure, Order not done: Patient Refusal" using "Patient Refusal SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.93)"
5. "Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
6. "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"
7. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
8. "Substance, Order: Autologous Blood Product" using "Autologous Blood Product SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.36)"
9. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
10. Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

Initial Population and Denominator Data Elements

Data elements 1- "Encounter, Performed: Encounter Inpatient," 2- “Laboratory Test, Performed; Hemoglobin blood serum plasma” and 6- “"Procedure, Performed: Selected Elective Surgical Procedures” are used to define the initial population and denominator of this measure.

On the feasibility scorecard, hospitals rated these data elements 1 and 6 as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year timeframe outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Four out of five hospitals rated capture of data element 6 as feasible or highly feasible, represented as a score of 2 or 3 out of 3. Facilities rating the data element as a 2 cited variation in clinical workflow and adoption of new technology as reasons for the lower rating. One site rated current state feasibility as a 1, as it did not currently have an interface between the OR scheduling system where this information was captured and the certified EHR technology. This site had plans to transition to an interfaced OR module in 1-2 years. All site rated the future state as highly feasible.

Hospitals reported low feasibility for workflow and data availability for data element 4. This data element must be recorded 45 to 14 days prior to the elective surgical procedure. While lab results are routinely captured as structured data, limited interoperability between hospitals and their community partners, such as clinics and lab centers, limits the availability of structured data for lab results occurring prior to the hospital encounter. Hospitals noted that many external results are received via fax, or as an electronic document, rather than in a format that can be structured and encoded in the EHR.

The Joint Commission views the Approval for Trial Use status as an opportunity to work with developers and implementers to further test and explore methods to improve feasibility for this data element.

Please refer to Appendix A for further findings related to external hemoglobin results.

Numerator Data Element

Data element 5- "Procedure, Performed: Blood Transfusion Administration" is used to define the numerator for this measure. Four out of five sites rated data element 5 as highly feasible, represented as a score of 3 out of 3. One hospital rated current feasibility as a 1 as blood transfusion was documented on a paper record, but had plans to implement blood transfusion via the EHR within six months.

Denominator Exclusions Data Elements

Data elements 3, 4, 7, 8, 9, and 10 are used to represent denominator exclusions.

Data elements 9- "Diagnosis: Traumatic Injury," and 10- "Diagnosis: Sickle Cell Disease and Related Blood Disorders" both represent encounter diagnoses. All hospitals rated these data elements as highly feasible. Discussion around these data elements suggested that while missing data may occur due to clinician practice related to updating the patient problem list, the functionality to support collection of this data element is well established.

Data element 3 represents blood transfusion, and data element 4 represents patient refusal. Both must be present in a record for a patient to meet the exclusion for patient refusal of blood transfusion. Four out of five sites rated data element 4 as highly feasible, represented as a score of 3 out of 3. Organizations reported greater variability, and lower feasibility, for capturing patient refusal of blood transfusion as structured data. Limitation to feasibility primarily stemmed from paper-based blood transfusion consent process and the limited support EHRs provide in general for consent. However, all sites discussed options for adding this documentation to their EHR, and felt that capture of refusal would be feasible in the future state.

Please refer to Appendix B for further findings related to patient refusal of transfusion

Feasibility for data element 7- "Procedure, Performed: Solid Organ Transplant" was found to be comparable to data element 6- "Procedure, Performed: Selected Elective Surgical Procedures." These data elements are found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform organ transplant, which would not use this data element.

Finally, all hospitals who use autologous blood products found data element 8- "Substance, Order: Autologous Blood Product," to be highly feasible. One hospital did not use autologous blood, and thus would not qualify for this exclusion.

Conclusion

Hospitals completing the feasibility scorecard largely reported the data elements required to calculate this measure to be feasible or highly feasible in the current state, with the exception of the data elements representing external hemoglobin results and patient refusal of transfusion. While hospitals felt patient refusal of transfusion would be feasible in the future, capture of hemoglobin results will require improvements in interoperability or workarounds to support data collection. Approval for Trial Use status will support The Joint Commission's efforts to further test this measure.

Appendix A: Feasibility Findings for External Hemoglobin Results

Note: The results below reflect responses recorded in the Feasibility Scorecards for PBM01: Preoperative Anemia Screening. At the time of testing, the timing parameter for hemoglobin in PBM05 was <=45 days prior to the first transfusion. Subsequent discussions with the technical advisory panel resulted in a change to the timing parameter to <=45 days and >=14 days, in order to align with PBM01. Because the aligned data elements are identical, the data from PBM01 is shown below, and the PBM05 scorecards have not been altered, retaining the assessment hospitals provided for the original timing parameter.

		Workflow		Data Availability		Data Accuracy		Data Element Definition		Data Standard	
Site		Score	Comments	Score	Comments	Score	Comments	Score	Comments	Score	Comments
1	Current	3		3		3		3		3	
	Future (3-5 years)	3		3		3		3		3	
2	Current	2		2	Sometimes procedures are performed without a hgb, and not always captured 14-45 days in advance	2		**3**		**3**	
	Future (3-5 years)	3	.	3		3		**3**		**3**	
3	Current	2	Not currently captured within time frame specified.	1	Never captured in time frame, most are scanned documents as they are not performed at facility	3		3		3	
	Future (3-5 years)	2	External results may or may not interface	2		3		3		3	
4	Current	2	Some results are scanned in from outside systems	2	Some results are scanned	3	Captured data is accurate	3		3	Not reported for scanned results
	Future (3-5 years)	3	Have educational plans in place to improve practice	2	Expect care model and technology will improve data availability, but still expect data will not always be available	3		3		3	
5	Current	2		2		3		3		3	
	Future (3-5 years)	2		2		3		3		3	

Appendix B: Feasibility Scorecard Findings for Patient Refusal of Blood Transfusion

Site		Workflow		Data Availability		Data Accuracy		Data Element Definition		Data Standard	
		Score	Comments	Score	Comments	Score	Comments	Score	Comments	Score	Comments
1	Current	1	Two types of refusal- refused in the moment, or refused across the board	1	Overall consent on chart- notes only in EHR	1		1		1	
	Future (3-5 years)	2		3	Could capture as a precaution order in the EHR	3		3		3	
2	Current	2	Blood bank has an alert, entered by blood bank staff	1		2		1		1	
	Future (3-5 years)	3	Are discussing having a "Bloodless Medicine" order that could trigger an alert within each encounter.	3		3		3		3	
3	Current	1		1		2		1		1	
	Future (3-5 years)	3	Capture in ICD-10	3		3		3		3	
4	Current	2		1		3		3		1	
	Future (3-5 years)	2	Uncertain- depends on how it would be defined in the EHR. Ideally, would have electronic signature. Would question accuracy of a flowsheet row.	2	If we had electronic signature, would be accurate. (Is electronic signature a priority? Unknown)	3		3		3	
5	Current	2	Using FYI functionality in Epic because it crosses encounters. This is a category list and currently you cannot attach a code. Currently also captured on paper on consent as well as the initial agreement.	1		3		3		1	Not able to capture in EHR product

	Future (3-5 years)	3		2	Dependent on vendor development	3		3		2	Dependent on vendor development

MEASURE WORKSHEET

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

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3032
Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: The STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score measures surgical performance for MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patent Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF). To assess overall quality, the STS MVRR +CABG Composite Score comprises two domains consisting of six measures:
Domain 1 – Absence of Operative Mortality
Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.
Domain 2 – Absence of Major Morbidity
Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Outcome data are collected on all patients and from all participants. For optimal measure reliability, participants meeting a volume threshold of at least 25 cases over 3 years receive a score for each of the two domains, plus an overall composite score. The overall composite score is created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:
1 star – lower-than-expected performance
2 stars – as-expected performance
3 stars – higher-than-expected performance

Developer Rationale: Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, mortality alone fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events. This composite provides a more comprehensive measure of overall quality and is timely due to the fact that MVRR+CABG comprises an increasing proportion of cardiac surgical practice and mortality risk is higher than for isolated MVRR [1-5].

References

1. Rankin JS, Feneley MP, Hickey M StJ, et al. A clinical comparison of mitral valve repair versus valve replacement in ischemic mitral regurgitation. J Thorac Cardiovasc Surg 95:165 77, 1988.

2. Glower DD, Tuttle RH, Shaw LK, et al: Patient survival characteristics after routine mitral valve repair for ischemic mitral regurgitation. J Thorac Cardiovasc Surg: 2005;129:860-868.
3. Milano CA, Daneshmand MA, Rankin JS, et al. Survival prognosis and surgical management of ischemic mitral regurgitation. Ann Thorac Surg 2008;86:735-744.
4. Daneshmand MA, Milano CA, Rankin JS, et al. Mitral valve repair for degenerative disease: A 20-year experience. Ann Thorac Surg 2009;88:1828-1837.
5. Daneshmand MA, Milano CA, Rankin JS, et al. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg 2010;90:1479-1486.

Numerator Statement: See Appendix
Denominator Statement: See Appendix
Denominator Exclusions: See Appendix

Measure Type: Composite
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Group/Practice, Facility

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

- This new composite measure of healthcare outcomes is comprised of absence of an operative mortality domain and an absence of major morbidity that includes any one or more of the identified complications.
- It is based on 7 NQF-endorsed measures of which 2 are mortality measures and 5 are cardiac surgery-related major morbidities.
  - 0122: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
  - 1502: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery
  - 0114: Risk-Adjusted Postoperative Renal Failure
  - 0115: Risk-Adjusted Surgical Re-exploration
  - 0129: Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
  - 0130: Risk-Adjusted Deep Sternal Wound Infection
  - 0131: Risk-Adjusted Stroke/Cerebrovascular Accident
- NQF criteria indicate that each component in a composite must meet the evidence sub criterion to justify its inclusion in the composite and be NQF-endorsed or evaluated as meeting measure evaluation criteria.
- The components of this composite are outcomes for which the required evidence is identification of a relationship between the outcome and at least one healthcare action that could achieve change in measure results. Information regarding service and/or care to impact mortality and 4 of the 5 morbidities is provided.
- The developer cites references that indicate that by taking morbidity into account the composite provides a more comprehensive measure of overall quality and is timely due to the fact that mitral valve repair/replacement plus CABG (MVRR+CABG) comprises an increasing proportion of cardiac surgical practice and mortality risk is higher than for isolated MVRR.
- Additional citations that address operative morbidity and mortality dating from 1998 through 2014 are provided.
- Approach to the work (see appendix S.14 and S.15)that underpins the measure is described and references provided.

Question for the Committee:

- *Is the information regarding modification and application of the model to the development of the composite clear and compelling?*
- *Does the Committee agree that the components together convey an appropriate measure of mitral valve surgery + CABG quality?*
- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm: Assess performance on outcome (Box 1) – Relationship between outcome and healthcare action (Box 2)

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

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

- The measure was calculated (see appendix 1b2) using STS Adult Cardiac Surgery Database data for patients undergoing MVRR + CABG in 2 consecutive overlapping 3 year time periods (July 2011- June 2-14 and July 2012 – 2015.
  - The developer reports that 2.3% and 2.9% of STS participants with ≥ 25 cases in the respective time periods have lower than expected performance on the measure based on 95% Bayesian credible interval. In comparison, 2.0% and 1.5% of all participants have lower than expected performance.

Disparities (see appendix 1b4)

Logistic regression was used to study associations of race (black), ethnicity (Hispanic), and insurance status (among patients < and ≥ age 65) with operative mortality and major morbidity while adjusting for the measure's risk adjustment model covariates. Odds ratios with 95% confidence intervals and p-values are summarized.

	OR (95% CI)	**P**	**OR (95% CI)**	**P**
Insurance status among patients age ≥65				
Medicare without Medicaid/Commercial	(ref)		(ref)	
Medicare and Medicaid dual eligible	1.03 (0.89, 1.20)	0.7058	1.24 (0.97, 1.58)	0.0806
Medicare Commercial without Medicaid	0.97 (0.90, 1.05)	0.5013	1.09 (0.95, 1.25)	0.2377
Insurance Status among patients age <65				
Commercial or HMO without Medicare/Medicaid	(ref)		(ref)	
Medicare or Medicaid	1.18 (1.06, 1.32)	0.0030	1.30 (1.04, 1.64)	0.0235
None/Self Pay	1.04 (0.89, 1.21)	0.6360	1.30 (0.94, 1.81)	0.1107
Other	1.11 (0.88, 1.39)	0.3893	1.15 (0.74, 1.79)	0.5381
Black race	1.19 (1.07, 1.33)	0.0015	0.86 (0.71, 1.04)	0.1301
Hispanic Ethnicity	1.15 (0.99, 1.34)	0.0712	0.74 (0.57, 0.97)	0.0316

Questions for the Committee:

- *In considering whether there is a gap in care that warrants a national performance measure, does the fact that each component of the measure represents occurrence of a serious adverse (never) event influence Committee thinking?*
- *Is handling of disparities data clear and are you aware of evidence that other disparities exist in this area of healthcare?*

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

1c. Composite - Quality Construct and Rationale
Maintenance measures – same emphasis on quality construct and rationale as for new measures.

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

- The development of the measure, the model upon which it is based and the modifications to the model to adjust for case mix are discussed in detail in the measure submission.
- The measure is based on a combination of the NQF-endorsed risk-adjusted operative mortality outcome measure and the risk-adjusted occurrence of any of five major complications. The developer states that an NQF-endorsed structure measure, database participation (0113), is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive composite scores. The developer notes that the MVRR + CABG Composite Score differs from the NQF-endorsed STS CABG Composite Score in that it does not include process measures. In this way it is similar to the NQF-endorsed STS AVR and AVR + CABG measures.
- The composite comprises 2 domains.
  - Domain 1 includes the proportion of patients (risk-adjusted) who do not experience operative mortality (death before hospital discharge or within 30 days of operation).
  - Domain 2 includes proportion of patients (risk-adjusted) who do not experience any major morbidity (occurrence of any one or more of prolonged ventilation, deep sternal wound infection, permanent stroke, renal failure, and reoperations for bleeding, prosthetic or native valve dysfunction, and other cardiac reasons, but not for non-cardiac reasons).
  - Participants receive a score for each of the 2 domains plus an overall composite score.
- The developer states that average mortality rates for the procedures of interest are at very low levels making differentiating performance based on mortality alone difficult in that it fails to take into account the fact that not all operative survivors received equal quality care. The composite, then, provides a more comprehensive measure of overall quality.
- The mortality domain corresponds to a single measure. The study endpoint for the morbidity domain combines multiple measures. Mortality rates were converted to survival rates and morbidity rates were converted to "absence of morbidity" rates. The developer notes that defining scores in this manner ensures that increasingly positive values reflect better performance. The overall composite score is created by "rolling up" the domain scores into a single number.
- Aggregation and weighting. To form the composite, the morbidity and mortality domains were rescaled by "dividing by their respective standard deviations across STS participants" and then adding the two domains together. After rescaling, the relative weights of risk-standardized mortality and risk-standardized major morbidity were 0.74 and 0.26, respectively. The weighting was assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the 2 domains.
- The minimum threshold for receiving a site-specific STS MVRR + CABG score (25 cases over 3 years) was selected on the basis of statistical testing – reliability =0.50.

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*
- *Does the Committee agree that the adaptation of NQF-endorsed measures meets the expectation that individual components of a composite meet NQF criteria?*

Preliminary rating for composite quality construct and rationale:

☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Note: Qualifies for high rating if Committee agrees that NQF expectation regarding endorsement or evaluation of component measures is satisfied.

Committee pre-evaluation comments

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

1a.

- composite health outcome connected to mult potential clinical actions
- developer reports that 2.3% and 2.9% of STS participants with > 25 cases in the respective time periods have

lower than expected performance on the measure based on 95% Bayesian credible interval. In comparison, 2.0% and 1.5% of all participants have lower than expected performance. ??

- usual issues with very small numbers -- hard to discriminate
- I like the move to a composite, though.
- 2% low outliers, 6% high outliers
- Cost of participation underestimates the requirement for a nurse abstractor in each facility.

1b.

- Gap exists

1c.

- Well constructed

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications
Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The data source for the measure is the STS Adult Cardiac Surgery Database. Data collection occurs through an electronic system using a detailed collection tool.
- The measure is specified for analysis at the group/practice level and is intended for use at the hospital/acute care setting.
- The composite score provides surgical performance for MVRR + CABG with or without concomitant ASD and PFO closures, TVr, or surgical ablation for AF. It is comprised of two domains: absence of operative mortality and absence of major morbidity (described under Construct). The measure is based on a combination of risk-adjusted mortality for MVRR + CABG and risk-adjusted occurrence of any one or more of 5 major complications.
- To adjust for case mix, logistic regression models for operative mortality and major complications were estimated using covariates from published STS 2008 risk models (see appendix S14 and S15). The developer notes that the main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population that did not meet inclusion/exclusion criteria for the 2008 model. The developer also noted that there is no existing model for predicting major morbidity as defined in the current study and that in future revisions of STS risk models, the composite measure will be calculated with the most up-to-date STS risk model for the MVRR + CABG population.
- As noted under Construct, the minimum threshold for receiving a site-specific STS MVRR + CABG composite score is 25 cases over 3 years.

Questions for the Committee :

Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

- The measure was developed and tested using data from the STS Adult Cardiac Surgery Database for 703 database participants for patients undergoing MVRR + CABG during July 2011 – June 2014.
- Risk model discrimination and calibration was assessed using data from 26,355 eligible patients during July 2011 – June 2014 from 1,038 database participants.
- The developer notes that to ensure adequate statistical precision, composite scores only for participants with at least 25 eligible cases during the 3 year measurement window will be reported.
- Estimated reliability of the measure using 3 years of data from participants with at least 25 total cases was 0.50 (95% CrI, 0.44 to 0.57). Results of comparisons using the same 3 years of data by including participants and (estimated reliability) with all cases (0.42), at least 25 cases (0.50) and at least 50 cases (0.62) is detailed. The developer notes that selecting the higher reliability would reduce the number of programs eligible to receive a score. For example, as reported, reliability of 0.62 would reduce the number of eligible programs from 341 to 143.
- The mathematical approach to signal-to-noise estimation is detailed.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the rationale for selection of the threshold for reporting clear and acceptable?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) – Empirical testing (Box 2) – Testing with measure score (Box 4) – Testing method described (Box 5) – Confidence that scores are reliable (Box 6)

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

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Describe any updates to validity testing: N/A

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

- Using the concept of performance categories, risk-adjusted mortality and morbidity rates were compared across 3 performance groups.
- The extent to which a participant's composite score remained stable across 2 consecutive reporting periods was assessed.
- Degree of uncertainty around a database participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CIs).

- Point estimates and CIs for an individual participant are reported with a comparison to benchmarks (overall average STS composite score and several percentiles) based on the national sample.
- Also, the composite measure result is converted into groups or categories using a Bayesian CI. If the CI around the composite score overlaps the overall STS average, the participant is performing at the as-expected level (indistinguishable from the average). If the CI is entirely above the national average, the participant has "higher-than-expected" performance and if entirely below the STS national average, the participants has "lower-than-expected" performance.

Validity testing results:

- Performance among database participants with at least 25 cases over 3 years during 2 time periods (July 2011 – June 2014 and July 2012 – June 2015 is provided:
- 310 (90.9%) and 314 (91.0%) performed as-expected during the 2 respective time periods;
- 8 (2.3%) and 10 (2.9%) had lower-than-expected performance;
- 23 (6.7%) and 21 (6.1%) had higher-than-expected performance.
- Also, performance data for all participants, including those with less than 25 cases, is provided.
- Using 2012 – 2015 data, risk-adjusted mortality and morbidity rates across the 3 performance categories based on 2011 – 2014 data were compared. Results were:
- 6.5% mortality and 29.7% morbidity among participants with as-expected performance;
- 11.1% mortality and 47.6% morbidity among participants with lower-than-expected performance;
- 4.3% mortality and 19.8% morbidity among participants with higher-than-expected performance.
- Developer interpretation of the results is that the composite measure behaves as expected and results are reasonably consistent across 2 consecutive overlapping time periods.

Questions for the Committee:

- *Is the work to demonstrate validity clear and complete such that you can agree that the score from this measure as specified is an indicator of quality?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

There are no exclusions.

o

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☒ **Yes** ☐ **No**

- Associations of race, ethnicity and insurance (see appendix 1b4) status with operative mortality and major morbidity were studied using logistic regression and results provided. Race (black) and ethnicity (Hispanic) are included in the STS 2008 models.
- As noted in the reliability section, logistic regression models for operative mortality and major complications were estimated using covariates from published STS 2008 risk models. The developer notes that the main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population that did not meet inclusion/exclusion criteria for the 2008 model.
- Covariates for the modified operative mortality model were identical to the STS 2008 operative mortality model and covariates for the new major morbidity model were identical to the STS 2008 operative mortality or major morbidity model except:
  - Adjustment variable for concomitant tricuspid repair was not in the 2008 model so was included;
  - Adjustment for tricuspid insufficiency using redefined categories of none or mild, moderate, and severe;
  - Adjustment for infectious endocarditis to provide for a treated category not included in the 2008 model.

- Estimated odds ratios from the modified STS 2008 models are provided.
- Discrimination and calibration was assessed using data from 26,355 patients undergoing MVRR + CABG during July 2011 – June 2014.
- Discrimination was gauged by calculating C-statistics for both models. Bootstrapping was used to estimate and adjust for "optimism" from estimating and evaluating the model on the same sample. Bootstrap-adjusted estimated C-statistic was 0.708 for morbidity model and 0.738 for mortality model. The developer notes that the numbers are comparable to the STS 2008 models when evaluated using the same sample (0.707 for morbidity and 0.738 for mortality)
- Calibration was evaluated using 5-fold cross validation. The approach is described in some detail. Expected to observed plots across the 5 samples are provided.
- The stated conclusion is that the risk models are well calibrated and have good discrimination power.

Questions for the Committee:

- *Does the approach and outcomes of modification of the STS 2008 model demonstrate an appropriate risk-adjustment strategy for the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Are there SDS factors that should be considered for evaluation as the measure is implemented?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- As noted in the validity testing section, performance among database participants with at least 25 cases over 3 years during 2 time periods (July 2011 – June 2014 and July 2012 – June 2015) demonstrate differences among 3 groupings:
  - 310 (90.9%) and 314 (91.0%) performed as-expected during the 2 respective time periods;
  - 8 (2.3%) and 10 (2.9%) had lower-than-expected performance;
  - 23 (6.7%) and 21 (6.1%) had higher-than-expected performance.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed. Single data source

2b7. Missing Data

- The overall frequency of missing data is reported at 0.55% for operative mortality and 0.44% for major complications.
  - Median participant-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications.
  - Percent of participants with >10% missing data was 0.7% for mortality and 1.5% for major complications.
  - As a sensitivity analysis, participant-specific mortality and complication rates were recalculated after excluding records with missing data from the denominator. There was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.
- The developer concludes that handling of missing outcome data is unlikely to impact performance results for the majority of participants.

2d. Composite measure: construction

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component

measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

- Correlation between each domain-specific estimate and overall composite score was calculated. Pearson correlations were 0.60 for mortality versus overall composite measure and 0.91 for morbidity domain score versus overall score.
- To form the composite, morbidity and mortality domains were rescaled by dividing by their respective standard deviations across STS participants and then adding the two domains together. The weighting was assessed by an expert panel to determine if an appropriate reflection of the relative importance of the 2 domains was provided. Relative weights in the final composite of risk-standardized mortality and risk standardized major morbidity were 0.74 and 0.26 respectively. This weighting was consistent with the expert panel's clinical assessment of each domain's relative importance.

Questions for the Committee:

- *Do the component measures fit the quality construct?*
- *Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

Guidance from the Validity Algorithm : Measure specification consistent with evidence (Box 1) – Potential threats to validity (Box 2) – Empirical validity testing (Box 3) – Face validity testing (Box 4) – Agreement that score can be used to distinguish quality (Box 5) – Moderate

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

Committee pre-evaluation comments

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

2a1.
- specifics well contstructed & retrievable from STS DB

2a2.
- reliable

2b1.
- Consist with evidence

2b2.
- Valid

2b3.
- Min miss data

2d.
- composite reflects quality in part (2 domains of mortal & morbid) & globally (star rating)

Criterion 3. Feasibility

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

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

- The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.
- Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants (representing over 90% of programs that provide cardiac surgery in US). Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.
- There are no additional costs for data collection specific to the measure. Costs to develop and maintain the

measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.
- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

Questions for the Committee:
Is the effort and cost associated with abstracting the required data elements appropriate to the value of the measure?

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

Committee pre-evaluation comments
Criteria 3: Feasibility

- Feasible

Criterion 4: Usability and Use

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

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

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

This new composite measure was developed in 2015 and will be published in 2016. STS plans to distribute participant-specific results in 2016 and begin public reporting "within the next year or so".

Questions for the Committee*:*
- *Does the Committee have concern about any potential unintended consequences?*

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

Committee pre-evaluation comments
Criteria 4: Usability and Use

- Not presently publically reported but planned

Criterion 5: Related and Competing Measures

Related or competing measures
- Related measures include STS measures that have been included in development of the composite or are otherwise related. They are harmonized.

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + CABG Composite Score
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

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

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

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

Episodes of Care; AQA Principles of Efficiency Measures).

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

Outcome

☒ Health outcome: 1. Operative Mortality; 2. Postoperative Major Morbidity

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

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

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

☐ Process:

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Operative Mortality

The incidence of mitral valve incompetence as a result of or coexistent with coronary artery disease is increasing as a result of a progressively older population of patients and recommendations for earlier surgical intervention among other factors. Patients undergoing combined coronary bypass and mitral valve replacement, in addition, increasingly have a larger number and severity of co-morbid risk factors. As a result, these patients have among one of the highest mortality rates of all surgical procedures. Mortality is likely the single most important negative outcome that can be associated with a surgical procedure. Critical evaluation of operative mortality allows one to evaluate the risk associated with a given procedure for various patient characteristics, and more importantly, aggressively search for ways to minimize that risk.

Major Morbidity

- Surgical re-exploration for bleeding remains a known complication following cardiac surgery. The literature documents that bleeding following coronary artery bypass surgery confers greater ICU stay and therefore greater resource consumption. It remains unknown and controversial whether long-term outcomes are worse for the isolated re-exploration for bleeding patients. However, Hein documents that patients with ICU stay > 3 days (with bleeding as multivariate risk factor for this outcome), have a long-term survival which is inferior to patients with ICU stay < 3 days. The patient consequences of this complication relates to the physiological stress of facing another operation and receiving blood products.
- A devastating complication of cardiac surgery is deep sternal wound infection. Patients with deep sternal wound infection require multiple surgeries to clear the infection, have longer hospital stays, greatly increased costs and increased early and late mortality. Care processes that influence the incidence of sternal infection span the first 3 major phases of surgical care. In the preoperative phase, routine patient decontamination and identification of active infections are key elements. In the intraoperative phase, impeccable surgical aseptic technique and timing of antibiotic administration are key processes. In the postoperative phase, aseptic wound care and detection of early wound inflammation are important in preventing delayed contamination and subsequent infection.
- Prolonged ventilation has been shown to substantially increase length of stay, the costs of care, and is associated with higher rates of respiratory failure, stroke, renal failure, and death. Modalities to decrease the rate of prolonged intubation include physician supervised protocols for extubation implemented by nurses and respiratory therapists, improved preoperative preparation of patients, reduction of postoperative bleeding, and intra-operative protocolized anesthesia care. Current implementation is highly variable and great opportunities to increase the implementation of evidence based care exist. Cardiac surgery programs with high implementation have lower than average rates of prolonged ventilation and significantly lower rates of adverse events.
- Postoperative renal failure is an occasional but serious complication in the cardiac surgical population and is a major determinant of short- and long-term survival. Identification of clinical precursors of postoperative renal insufficiency and improvement in perioperative treatment of this high-risk group will improve the long-term

survival of our patients. By implementing known recommendations (delay heart surgery, when possible, after cardiac catheterization, maintain mean CPB perfusion pressure at 80% of preop BP, etc), postoperative kidney injury should be significantly reduced.

- Postoperative stroke/CVA produces significant short- and long-term often devastating effects to patients and their families. It is associated with significant increases in death, respiratory failure, renal failure, length of stay, and cost of care. Modalities to decrease the rate of postoperative stroke include measures to maintain blood pressure and perfusion, glycemic control, avoidance of atrial fibrillation, anticoagulation protocols, etc. Many opportunities exist to decrease stroke rates by increasing implementation of evidence based strategies.

References – Operative Mortality

- Birkmeyer NJ, Marrin CA, et al. Decreasing mortality for aortic and mitral valve surgery in Northern New England. Northern New England Cardiovascular Disease Study Group. Ann Thorac Surg. 2000;70(2):432-437.
- Edwards FH, Peterson ED, et al. Prediction of operative mortality following valve replacement surgery. JACC. 37:3:885-892.
- Goodney PP, O'Connor GT, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? Ann Thorac Surg. 2003;76:1131-1137.
- Mehta RH, Eagle KA, et al. Influence of age on outcomes in patients undergoing mitral valve replacement. Ann Thorac Surg. 2002;74:1459-1467.
- Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, Normand SL, DeLong ER, Shewan CM, Dokholyan RS, Peterson ED, Edwards FH, Anderson RP. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul; 88(1 Suppl):S43-62.
- Miyata H, Motomura N, Tsukihara H, Takamoto S; Japan Cardiovascular Surgery Database. Risk models including high-risk cardiovascular procedures: clinical predictors of mortality and morbidity. Eur J Cardiothorac Surg. 2010 Nov 1
- Vassileva CM, Boley T, Markwell S, Hazelrigg S. Meta-analysis of short-term and long-term survival following repair versus replacement for ischemic mitral regurgitation. Eur J Cardiothorac Surg. 2010 Aug 18.
- Daneshmand MA, Milano CA, Rankin JS, Honeycutt EF, Shaw LK, Davis RD, Wolfe WG, Glower DD, Smith PK. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg. 2010 Nov; 90(5):1479-85
- Acker MA, Parides MK, Perrault LP et al (members of Cardiothoracic Surgical Trials Network). Mitral-valve repair versus replacement for severe ischemic mitral regurgitation. N Engl J Med 2014; 370:23-32

References – Major Morbidity

- Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg.* 2003;75:1856-1865.
- Hein OV, Birnbaum J, Wernecke K, England M, Knoertz W, Spies C. Prolonged Intensive Care Unit Stay in Cardiac Surgery: Risk Factors and Long-Term Survival. *Ann Thor Surg* 2006;81:880-85.
- Stamou SC, Camp SL, Stiegel RM, et al. Quality improvement program decreases mortality after cardiac surgery. *J Thorac Cardiovasc Surg* 2008;136:494-499.
- Braxton JH, Marrin CA, McGrath PD, et al. 10-Year follow-up of patients with and without mediastinitis. Semin Thorac Cardiovasc Surg. 2004;16:70–76.
- Graf K, Ott E, Vonberg RP, et al. Economic aspects of deep sternal wound infections. Eur J Cardiothorac Surg 2010;37:893-96.
- Edwards FH, Engelman RM, Houck P et al. The Society of Thoracic Surgeons Practice Guideline Series: Antibiotic Prophylaxis in Cardiac Surgery, Part I: Duration. Ann Thorac Surg 2006;81: 397-404,
- Wilson APL, Gibbons C, Reeves BC, et al. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. BMJ 2004;329:720-24.
- Filsoufi F, Castillo JG, Rahmanian PB, et al. Epidemiology of deep sternal wound infection in cardiac surgery. J Cardiothorac Vasc Anesth 2009;23:488-94.
- Koch CG, Nowicki ER, Rajeswaran J, et al. When the timing is right: antibiotic timing and infection after cardiac surgery. J Thorac Cardiovasc Surg 2012;144:931-37.

- Paul M, Raz, A, Leibovici L, et al. Sternal wound infection after coronary artery bypass graft surgery: validation of existing risk scores. J Thorac Cardiovasc Surg 2007;133:397-403.
- Lazar HL, Ketchedjian A, Haime M, et al. Topical Vancomycin in combination with perioperative antibiotics and tight glycemic control helps to eliminate sternal wound infections. J Thorac Cardiovasc Surg 2014;148:1035-40.
- Miyahara K, MatsuuraA, Takemura H, et al. Implementation of bundled interventions greatly decreases deep sternal wound infection following cardiovascular surgery. J Thorac Cardiovasc Surg 2014;148:2381-88.
- Matros E, Aranki, SF, Bayer LR, et al. Reduction in incidence of deep sternal wound infections: random or real? J Thorac Cardiovasc Surg 2010;139:680-85.
- Bardell T, Legare JF, Buth KJ, et al. ICU readmission after cardiac surgery. Eur J Cardiothorac Surg. 2003;23(3):354-359.
- Meade MO, Guyatt G, Butler R, et al. Trials comparing early vs late extubation following cardiovascular surgery. Chest. 2001:120(6 Suppl):445S-453S.
- Naughton C, Reilly N, Powroznyk A, et al. Factors determining the duration of tracheal intubation in cardiac surgery: a single-centre sequential patient audit. Eur J Anaesthesiol. 2003;20(3):225-233.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg. 2004;77:1137-1139.
- Knapik P, Ciesla D, Borowik D, Czempik P, Knapik T. Prolonged ventilation post cardiac surgery – tips and pitfalls of the prediction game. J Cardiothorac Surg 2011;6:158.
- Alsabbagh MM, Asmar A, Ejaz NI, Aiyer RK, Kambhampati G, Ejaz AA. Update on clinical trials for the prevention of acute kidney injury in patients undergoing cardiac surgery. Am J Surg 2013;206:86-95
- Arora P, Kolli, H, Nainani N, Nader N, Lohr J. Preventable risk factors for acute kidney injury in patients undergoing cardiac surgery. J Cardiothorac Vasc Anesth 2012; 26:687-697.
- Boldt J, Brenner T, Lehmann A, Suttner SW, Kumle B, Isgro F. Is kidney function altered by the duration of cardiopulmonary bypass? Ann Thorac Surg. 2003;75(3):906-912.
- Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. Am J Med. 1998;104(4):343-348
- Conlon PJ, Stafford-Smith M, White WD, Newman MF, King S, Winn MP, Landolfo K. Acute renal failure following cardiac surgery. Nephrol Dial Transplant. 1999;14(5):1158-1162.
- Haase M, Haase-Fielitz A, Bellomo R, Devarajan P, Story D, Matalanis G, Reade MC, Bagshaw SM, Seevanayagam N, Seevanayagam S, Doolan L, Buxton B, Dragun D. Sodium bicarbonate to prevent increases in serum creatinine after cardiac surgery: a pilot double-blind, randomized trial. Crit Care Ned 2009;37:39-47.
- Kramer RS, Quinn RD, Groom RC, Braxton JH, Malenka DJ, Kellett MA, Brown JR for the Northern New England Cardiovascular Disease Study Group. Same admission cardiac catheterization and cardiac surgery: is there an increased incidence of acute kidney injury? Ann Thorac Surg 2010;90:1418-1424.
- Mangano CM, Diamondstone LS, Ramsay JG, et al. Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization: the Multicenter Study of Perioperative Ischemia Research Group. Ann Intern Med. 1998;128(3):194-203.
- Ranucci M, Ballotta A, Agnelli B, Frigiola A, Mencanti L, Castelvecchio S, for the Surgical and Clinical Outcome Research (SCORE) Group. Acute kidney injury in patients undergoing cardiac surgery and coronary angiography on the same day. Ann Thorac Surg 2103;95:513-519.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, DeLong ER, O'Brien SM, Shewan CM, Dokholyan RS, Peterson ED. Quality Measurement in adult cardiac surgery: Part 1-conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12
- Tang AT, Alexiou C, Hsu J, Sheppard SV, Haw MP, Ohri SK. Leukodepletion reduces renal injury in coronary revascularization: a prospective randomized study. Ann Thorac Surg. 2002;74(2):372-327; discussion 377.
- Amory DW, Grigore A, Amory JK, et al. Neuroprotection is associated with beta-adrenergic receptor antagonists during cardiac surgery: evidence from 1,575 patients. J Cardiothorac Vasc Anesth. 2002;16(3):270-277.
- Arsenault KA, Yusus AM, Crystal E, Healey JS, Morillo CA, Nair GM et al. Interventions for preventing postoperative atrial fibrillation in patients undergoing heart surgery. Cocrane Database Syst Rev. 2013; 1:CD003611
- Bucerius J, Gummert JF, Borger MA, et al. Predictors of delirium after cardiac surgery delirium: effect of beating-heart (off-pump) surgery. J Thorac Cardiovasc Surg. 2004;127(1):57-64.

- Inoue K, Luth JU, Pottkamper D, et al. Incidence and risk factors of perioperative cerebral complications: heart transplantation compared to coronary artery bypass grafting and valve surgery. J Cardiovasc Surg. 1998;39(2):201-208.
- Rosenberger P, Shernan SK, Loffler M, Shekar PS, Fox JA, Tuli JK, Nowak M and Eltzschig HK. The influence of epiaortic ultrasonograpy n intraoperative surgical management in 6051 cardiac surgical patients. Ann Thorac Surg. 2008; 85: 548-53.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Please see response above.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

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

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

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

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

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

Evidence_Form.STS_MVRR-CABG_Composite_Score.docx

1b. Performance Gap

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

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

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

N/A

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

See Appendix.

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

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

See Appendix.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Please see attached evidence form for detailed information.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
Please see attached evidence form for list of references.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

1d. Composite Quality Construct and Rationale

1d.1. **A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.**

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
    - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
    - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

1d.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1d.2. **Describe the quality construct, including:**

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

The STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score measures surgical performance for MVRR + CABG with or without concomitant Atrial Septal Defect (ASD) and Patient Foramen Ovale (PFO) closures, tricuspid valve repair (TVr), or surgical ablation for atrial fibrillation (AF). Similar to other STS composite measures, this measure is based on a combination of the NQF-endorsed risk-adjusted operative mortality outcome measure and the risk-adjusted occurrence of any of five major complications. An NQF-endorsed structure measure, database participation, is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive composite scores. To assess overall quality, the composite comprises the following two domains:

Domain 1 – Absence of Operative Mortality
Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity
Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:
1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars.

Similar to the NQF-endorsed STS AVR and AVR+CABG measures, the MVRR+CABG Composite Score differs from the NQF-endorsed STS CABG Composite Score in that it does not include process measures. This reflects the fact that for MVRR+CABG, in comparison with isolated CABG surgery, no widely accepted process measures meeting performance metric criteria currently exist.

1d.3. **Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.**
Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, mortality alone fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events. This composite provides a more comprehensive measure of overall quality and is timely due to the fact that MVRR+CABG comprises an increasing proportion of cardiac surgical practice and mortality risk is higher than for isolated MVRR [1-5].

References
1. Rankin JS, Feneley MP, Hickey M StJ, et al. A clinical comparison of mitral valve repair versus valve replacement in ischemic mitral regurgitation. J Thorac Cardiovasc Surg 95:165 77, 1988.
2. Glower DD, Tuttle RH, Shaw LK, et al: Patient survival characteristics after routine mitral valve repair for ischemic mitral regurgitation. J Thorac Cardiovasc Surg: 2005;129:860-868.
3. Milano CA, Daneshmand MA, Rankin JS, et al. Survival prognosis and surgical management of ischemic mitral regurgitation. Ann Thorac Surg 2008;86:735-744.
4. Daneshmand MA, Milano CA, Rankin JS, et al. Mitral valve repair for degenerative disease: A 20-year experience. Ann Thorac Surg 2009;88:1828-1837.
5. Daneshmand MA, Milano CA, Rankin JS, et al. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg 2010;90:1479-1486.

1d.4. **Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.**
The mortality domain corresponds to a single measure, while the study endpoint for the morbidity domain combines multiple measures and thus is a composite endpoint. To enhance interpretation, mortality rates were converted to survival rates (risk-standardized survival rate = 100 – risk-standardized mortality rate), and morbidity rates were converted to "absence of morbidity" rates (risk-standardized absence of morbidity rate =100 – risk-standardized morbidity rate). Defining scores in this manner ensures that increasingly positive values reflect better performance. The overall composite score is created by "rolling up" the domain scores into a single number.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

evaluated against the remaining criteria.

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

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

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Complications, Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/STSAdultCVDataCollectionForm2_73_Annotated.pdf,
http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;Addition info in Comments Section

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

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
See Appendix

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
See Appendix

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

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

See Appendix

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
See Appendix

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See Appendix

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See Appendix

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
See Appendix

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please see discussion under section S.4 and attached manuscripts.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.55% of records for operative mortality and 0.44% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

The overall frequency of missing data was 0.55% for operative mortality and 0.44% for major complications. The median participant-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications. The percent of participants with >10% missing data was 0.7% for mortality and 1.5% for major complications. As a sensitivity analysis, we re-calculated participant-specific mortality and complication rates after excluding records with missing data from the denominator. As shown in the figures in section 2b7.2. of the testing attachment, there was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database – Version 2.73; STS Adult Cardiac Surgery Database Version 2.81 went live on July 1, 2014.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Please see section S.4

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Testing_Form.STS_MVRR-CABG_Composite_Score-636008081648083729.docx

NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2d)

Measure Number (*if previously endorsed*)**: 14T**
Composite Measure Title: STS Mitral Valve Repair/Replacement (MVRR) + CABG Composite Score
Date of Submission: 6/5/2016
Composite Construction:
☒Two or more individual performance measure scores combined into one score
☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)
☒ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

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

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

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

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

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

2b4. 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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

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

2d1. 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

2d2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

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

Notes

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

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

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

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. If different data sources are used for different components in the composite, indicate the component after the checkbox.*)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: 14T	☐ other: 14T

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
STS Adult Cardiac Surgery Database Version 2.73

1.3. What are the dates of the data used in testing?
July 2011 – June 2014

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 14T	☐ other: 14T

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the*

analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)

The measure was developed and tested using STS Adult Cardiac Surgery Database data from 703 participants for patients undergoing mitral valve repair/replacement (MVRR) + CABG during July 2011 – June 2014. Only participants with at least 10 eligible records during this period were included in the hierarchical model for estimating composite scores. The table below summarizes the distribution of participant-specific denominators (number of eligible patients) and participant-specific mortality and morbidity rates.

Stat	N (Denominator)	% Mortality	% Morbidity
N	703	703	703
Mean	35	6.8	31.7
STD	34	6.1	13.3
IQR	27	6.4	17.3
0%	10	0.0	0.0
10%	12	0.0	16.3
20%	14	0.0	20.3
30%	17	3.6	24.0
40%	20	4.8	27.0
50%	23	5.9	30.3
60%	28	7.1	33.3
70%	36	8.6	37.8
80%	50	10.5	42.1
90%	73	14.3	50.0
100%	503	38.9	76.9

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

For assessing risk model discrimination and calibration, the sample included eligible 26,355 patient operation records from 1,038 STS participants. For estimating composite scores, the sample was limited to participants with at least 10 eligible cases over the 3-year study period (24,740 patients, 703 centers).

The table below summarizes the baseline characteristics of patients who were included in the estimation of composite scores during July 2011 – June 2014.

Variable	Effects	Overall N=24,740	MV Repair N=16,300	MV Replace N=8,440
Patient Age	Median (IQR)	69.0 (62.0, 76.0)	69.0 (61.0, 76.0)	70.0 (62.0, 77.0)
Surgery Year	2011 (half year)	3,958 (16.0%)	2,664 (16.3%)	1,294 (15.3%)
	2012	8,230 (33.3%)	5,525 (33.9%)	2,705 (32.0%)
	2013	8,417 (34.0%)	5,571 (34.2%)	2,846 (33.7%)
	2014 (half year)	4,135 (16.7%)	2,540 (15.6%)	1,595 (18.9%)
Previous Cardiac Surgery	No	16,167 (65.3%)	11,047 (67.8%)	5,120 (60.7%)
	Yes	8,532 (34.5%)	5,228 (32.1%)	3,304 (39.1%)
	Missing	41 (0.2%)	25 (0.2%)	16 (0.2%)
Previous PCI	No	2,411 (28.3%)	1,240 (23.7%)	1,171 (35.4%)
	Yes	6,112 (71.6%)	3,987 (76.3%)	2,125 (64.3%)
	Missing	9 (0.1%)	1 (0.0%)	8 (0.2%)
Previous PCI - Interval	<=6 Hours	88 (1.4%)	39 (1.0%)	49 (2.3%)
	>6 Hours	6,009 (98.3%)	3,939 (98.8%)	2,070 (97.4%)
	Missing	15 (0.2%)	9 (0.2%)	6 (0.3%)
Ejection Fraction (%)	Median (IQR)	48.0 (35.0, 60.0)	45.0 (30.0, 56.0)	53.0 (40.0, 60.0)
	Missing	428 (1.7%)	255 (1.6%)	173 (2.0%)

LV End-Systolic Dimension	Median (IQR)	38.0 (31.3, 46.0)	39.0 (33.0, 47.2)	35.9 (29.8, 43.0)
	Missing	12,082 (48.8%)	7,798 (47.8%)	4,284 (50.8%)
LV End-Diastolic Dimension	Median (IQR)	53.0 (47.0, 59.0)	54.0 (48.0, 59.0)	51.0 (45.2, 57.0)
	Missing	12,060 (48.7%)	7,769 (47.7%)	4,291 (50.8%)
PA Systolic Pressure	Median (IQR)	43.0 (33.8, 55.0)	42.0 (32.0, 53.0)	46.0 (35.0, 59.0)
	Missing	36 (0.3%)	24 (0.3%)	12 (0.2%)
Aortic Valve Disease (mild)	No	17,294 (69.9%)	11,554 (70.9%)	5,740 (68.0%)
	Yes	7,340 (29.7%)	4,673 (28.7%)	2,667 (31.6%)
	Missing	106 (0.4%)	73 (0.4%)	33 (0.4%)
Aortic Valve Disease Etiol.	Degenerative	2,536 (34.6%)	1,544 (33.0%)	992 (37.2%)
	Endocarditis	19 (0.3%)	9 (0.2%)	10 (0.4%)
	Congenital	32 (0.4%)	24 (0.5%)	8 (0.3%)
	Rheumatic	74 (1.0%)	13 (0.3%)	61 (2.3%)
	Primary Aortic Dis.	12 (0.2%)	6 (0.1%)	6 (0.2%)
	LVOTO	8 (0.1%)	2 (0.0%)	6 (0.2%)
	Tumor	2 (0.0%)	1 (0.0%)	1 (0.0%)
	Trauma	1 (0.0%)	1 (0.0%)	0 (0.0%)
	Other	1,080 (14.7%)	707 (15.1%)	373 (14.0%)
	Missing	3,574 (48.7%)	2,366 (50.6%)	1,208 (45.3%)
Aortic Valve Stenosis	No	6,269 (85.4%)	4,090 (87.5%)	2,179 (81.7%)
	Yes	909 (12.4%)	465 (10.0%)	444 (16.6%)
	Missing	162 (2.2%)	118 (2.5%)	44 (1.6%)
Aortic Valve Insufficiency	None	541 (7.4%)	316 (6.8%)	225 (8.4%)
	Trivial	2,884 (39.3%)	1,917 (41.0%)	967 (36.3%)
	Mild	3,116 (42.5%)	1,964 (42.0%)	1,152 (43.2%)
	Moderate	742 (10.1%)	440 (9.4%)	302 (11.3%)
	Severe	40 (0.5%)	26 (0.6%)	14 (0.5%)
	Missing	17 (0.2%)	10 (0.2%)	7 (0.3%)
MV Disease Etiol.	Degenerative	12,807 (53.0%)	8,877 (55.9%)	3,930 (47.3%)
	Endocarditis	643 (2.7%)	174 (1.1%)	469 (5.6%)
	Rheumatic	1,129 (4.7%)	147 (0.9%)	982 (11.8%)
	Ischemic	3,300 (13.6%)	2,357 (14.8%)	943 (11.3%)
	Congenital	73 (0.3%)	56 (0.4%)	17 (0.2%)
	HOCM	39 (0.2%)	20 (0.1%)	19 (0.2%)
	Tumor	77 (0.3%)	49 (0.3%)	28 (0.3%)
	Trauma	5 (0.0%)	4 (0.0%)	1 (0.0%)
	Non-isch Myopathy	153 (0.6%)	103 (0.6%)	50 (0.6%)
	Other	2,028 (8.4%)	1,287 (8.1%)	741 (8.9%)
	Missing	3,930 (16.3%)	2,799 (17.6%)	1,131 (13.6%)
MV Degenerative Location	Posterior Leaflet	953 (34.7%)	780 (41.1%)	173 (20.4%)
	Anterior Leaflet	303 (11.0%)	178 (9.4%)	125 (14.7%)
	Bi-leaflet	445 (16.2%)	225 (11.9%)	220 (25.9%)
	Missing	1,043 (38.0%)	713 (37.6%)	330 (38.9%)
Mitral Annular Disease Type	Pure Ann Dilation	950 (39.7%)	872 (47.8%)	78 (13.8%)
	Ann Calcification	397 (16.6%)	238 (13.1%)	159 (28.0%)
	Missing	1,043 (43.6%)	713 (39.1%)	330 (58.2%)
MV Ischemic Type	Acute	1,105 (33.5%)	638 (27.1%)	467 (49.5%)
	Chronic	2,081 (63.1%)	1,630 (69.2%)	451 (47.8%)
	Missing	114 (3.5%)	89 (3.8%)	25 (2.7%)
MV NYHA Functional Class	I	4,528 (18.7%)	3,403 (21.4%)	1,125 (13.5%)
	II	5,429 (22.4%)	3,612 (22.8%)	1,817 (21.9%)
	IIIa	2,926 (12.1%)	1,509 (9.5%)	1,417 (17.0%)
	IIIb	2,222 (9.2%)	1,476 (9.3%)	746 (9.0%)
	Missing	9,079 (37.5%)	5,873 (37.0%)	3,206 (38.6%)
Mitral Stenosis	No	21,815 (90.2%)	15,174 (95.6%)	6,641 (79.9%)
	Yes	1,775 (7.3%)	274 (1.7%)	1,501 (18.1%)
	Missing	594 (2.5%)	425 (2.7%)	169 (2.0%)
Mitral Insufficiency	None	260 (1.1%)	87 (0.5%)	173 (2.1%)

	Trivial	242 (1.0%)	118 (0.7%)	124 (1.5%)
	Mild	1,070 (4.4%)	625 (3.9%)	445 (5.4%)
	Moderate	5,722 (23.7%)	4,374 (27.6%)	1,348 (16.2%)
	Severe	16,795 (69.4%)	10,606 (66.8%)	6,189 (74.5%)
	Missing	95 (0.4%)	63 (0.4%)	32 (0.4%)
Tricuspid Valve Disease	No	9,480 (38.3%)	6,266 (38.4%)	3,214 (38.1%)
	Yes	15,155 (61.3%)	9,961 (61.1%)	5,194 (61.5%)
	Missing	105 (0.4%)	73 (0.4%)	32 (0.4%)
Tricuspid Insufficiency	None	57 (0.4%)	36 (0.4%)	21 (0.4%)
	Trivial	3,159 (20.8%)	2,254 (22.6%)	905 (17.4%)
	Mild	6,052 (39.9%)	4,020 (40.4%)	2,032 (39.1%)
	Moderate	4,137 (27.3%)	2,603 (26.1%)	1,534 (29.5%)
	Severe	1,671 (11.0%)	1,000 (10.0%)	671 (12.9%)
	Missing	79 (0.5%)	48 (0.5%)	31 (0.6%)
	Missing	8 (0.1%)	4 (0.1%)	4 (0.2%)
Operative Approach	Full Sternotomy	24,560 (99.3%)	16,191 (99.3%)	8,369 (99.2%)
	Partial Sternotomy	71 (0.3%)	41 (0.3%)	30 (0.4%)
	Rt or Lt Parasternal	4 (0.0%)	2 (0.0%)	2 (0.0%)
	Left Thoracotomy	6 (0.0%)	1 (0.0%)	5 (0.1%)
	Right Thoracotomy	14 (0.1%)	7 (0.0%)	7 (0.1%)
	Transverse	1 (0.0%)	1 (0.0%)	0 (0.0%)
	Minimally Invasive	27 (0.1%)	24 (0.1%)	3 (0.0%)
	Missing	57 (0.2%)	33 (0.2%)	24 (0.3%)
Robotic Assisted	No	24,472 (98.9%)	16,116 (98.9%)	8,356 (99.0%)
	Yes	39 (0.2%)	26 (0.2%)	13 (0.2%)
	Missing	229 (0.9%)	158 (1.0%)	71 (0.8%)
Mitral Valve Procedure	Repair	16,300 (65.9%)	16,300 (100.0%)	0 (0.0%)
	Replacement	8,440 (34.1%)	0 (0.0%)	8,440 (100.0%)
MV Repair - Annuloplasty	No	729 (4.5%)	729 (4.5%)	. (.%)
	Yes	15,434 (94.7%)	15,434 (94.7%)	. (.%)
	Missing	137 (0.8%)	137 (0.8%)	. (.%)
MV Repair - Leaf Resection	No	13,241 (81.2%)	13,241 (81.2%)	. (.%)
	Yes	2,660 (16.3%)	2,660 (16.3%)	. (.%)
	Missing	399 (2.4%)	399 (2.4%)	. (.%)
MV Leaflet Resection Type	Triangular	1,323 (49.7%)	1,323 (49.7%)	. (.%)
	Quadrangular	991 (37.3%)	991 (37.3%)	. (.%)
	Other	255 (9.6%)	255 (9.6%)	. (.%)
	Missing	91 (3.4%)	91 (3.4%)	. (.%)
MV Repair Location	Anterior	181 (6.8%)	181 (6.8%)	. (.%)
	Posterior	2,279 (85.7%)	2,279 (85.7%)	. (.%)
	Bileaflet	143 (5.4%)	143 (5.4%)	. (.%)
	Missing	57 (2.1%)	57 (2.1%)	. (.%)
MV Sliding Plasty	No	15,290 (93.8%)	15,290 (93.8%)	. (.%)
	Yes	557 (3.4%)	557 (3.4%)	. (.%)
	Missing	453 (2.8%)	453 (2.8%)	. (.%)
MV Annular Decalcification	No	15,740 (96.6%)	15,740 (96.6%)	. (.%)
	Yes	136 (0.8%)	136 (0.8%)	. (.%)
	Missing	424 (2.6%)	424 (2.6%)	. (.%)
PTFE Chordal Replacement	No	14,529 (89.1%)	14,529 (89.1%)	. (.%)
	Yes	1,316 (8.1%)	1,316 (8.1%)	. (.%)
	Missing	455 (2.8%)	455 (2.8%)	. (.%)
Neo-chordal Number	Median (IQR)	2.0 (1.0, 3.0)	2.0 (1.0, 3.0)	. (., .)
	Missing	39 (3.0%)	39 (3.0%)	. (.%)
MV Chordal Transfer	0	15,594 (95.7%)	15,594 (95.7%)	. (.%)
	Yes	263 (1.6%)	263 (1.6%)	. (.%)
	Missing	443 (2.7%)	443 (2.7%)	. (.%)
MV Leaflet Patch	0	15,641 (96.0%)	15,641 (96.0%)	. (.%)
	Yes	215 (1.3%)	215 (1.3%)	. (.%)
	Missing	444 (2.7%)	444 (2.7%)	. (.%)
MV Edge To Edge Repair	No	15,189 (93.2%)	15,189 (93.2%)	. (.%)

	Yes	640 (3.9%)	640 (3.9%)	. (.%)
	Missing	471 (2.9%)	471 (2.9%)	. (.%)
Mitral Commissurotomy	0	667 (4.1%)	667 (4.1%)	. (.%)
	Yes	15,633 (95.9%)	15,633 (95.9%)	. (.%)
MV Repair Attempt	No	6,989 (82.8%)	. (.%)	6,989 (82.8%)
	Yes	1,260 (14.9%)	. (.%)	1,260 (14.9%)
	Missing	191 (2.3%)	. (.%)	191 (2.3%)
MV Replace - Chordal Pres.	None	3,330 (13.5%)	1,651 (10.1%)	1,679 (19.9%)
	Anterior	607 (2.5%)	344 (2.1%)	263 (3.1%)
	Posterior	1,938 (7.8%)	186 (1.1%)	1,752 (20.8%)
	Both	12,512 (50.6%)	8,865 (54.4%)	3,647 (43.2%)
	Missing	6,353 (25.7%)	5,254 (32.2%)	1,099 (13.0%)
	Missing	847 (3.4%)	549 (3.4%)	298 (3.5%)
Postop TEE MR Grade	None	10,633 (55.2%)	6,683 (51.2%)	3,950 (63.6%)
	Trace/Trivial	4,718 (24.5%)	3,548 (27.2%)	1,170 (18.8%)
	Mild	1,653 (8.6%)	1,399 (10.7%)	254 (4.1%)
	Moderate	682 (3.5%)	538 (4.1%)	144 (2.3%)
	Severe	662 (3.4%)	395 (3.0%)	267 (4.3%)
	Missing	918 (4.8%)	494 (3.8%)	424 (6.8%)
Operative Mortality	No	23,208 (93.8%)	15,500 (95.1%)	7,708 (91.3%)
	Yes	1,532 (6.2%)	800 (4.9%)	732 (8.7%)
Any Post-Op Events	No	9,213 (37.2%)	6,531 (40.1%)	2,682 (31.8%)
	Yes	15,475 (62.6%)	9,736 (59.7%)	5,739 (68.0%)
	Missing	52 (0.2%)	33 (0.2%)	19 (0.2%)
Reop - Bleeding	No	23,683 (95.7%)	15,679 (96.2%)	8,004 (94.8%)
	Yes	984 (4.0%)	575 (3.5%)	409 (4.8%)
	Missing	73 (0.3%)	46 (0.3%)	27 (0.3%)
Reop - Valve Dysfunction	No	24,621 (99.5%)	16,222 (99.5%)	8,399 (99.5%)
	Yes	44 (0.2%)	30 (0.2%)	14 (0.2%)
	Missing	75 (0.3%)	48 (0.3%)	27 (0.3%)
Reop - Other Cardiac	No	24,300 (98.2%)	16,028 (98.3%)	8,272 (98.0%)
	Yes	364 (1.5%)	225 (1.4%)	139 (1.6%)
	Missing	76 (0.3%)	47 (0.3%)	29 (0.3%)
Reop - Other Non Cardiac	No	23,570 (95.3%)	15,612 (95.8%)	7,958 (94.3%)
	Yes	1,094 (4.4%)	639 (3.9%)	455 (5.4%)
	Missing	76 (0.3%)	49 (0.3%)	27 (0.3%)
Deep Sternal Infection	No	24,559 (99.3%)	16,191 (99.3%)	8,368 (99.1%)
	Yes	102 (0.4%)	60 (0.4%)	42 (0.5%)
	Missing	79 (0.3%)	49 (0.3%)	30 (0.4%)
Permanent Stroke	No	23,975 (96.9%)	15,824 (97.1%)	8,151 (96.6%)
	Yes	684 (2.8%)	423 (2.6%)	261 (3.1%)
	Missing	81 (0.3%)	53 (0.3%)	28 (0.3%)
Prolonged Ventilation	No	18,216 (73.6%)	12,499 (76.7%)	5,717 (67.7%)
	Yes	6,428 (26.0%)	3,731 (22.9%)	2,697 (32.0%)
	Missing	96 (0.4%)	70 (0.4%)	26 (0.3%)
Renal Failure	No	23,021 (93.1%)	15,320 (94.0%)	7,701 (91.2%)
	Yes	1,647 (6.7%)	934 (5.7%)	713 (8.4%)
	Missing	72 (0.3%)	46 (0.3%)	26 (0.3%)

PCI=percutaneous coronary intervention, LV=left ventricular, PA=pulmonary artery, Etiol.=etiology, MV=mitral valve, NYHA=New York Heart Association, PTFE=polytetrafluoroethylene, Pres=preservation, TEE=transesophageal echo, Reop=reoperation

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

For developing and evaluating case mix adjustment procedures, we used data from 26,355 eligible patients at 1,038 participants undergoing MVRR+CABG during July 2011 – June 2014.

For estimating participant-specific composite scores, the analysis was restricted to data from STS participants with at least 10 eligible cases during July 2011 – June 2014 (N = 24,740 patient records, 703 participants).

For assessing the consistency of results over time, we re-estimated composite scores using data from July 2012 – June 2015 (N = 24,376 patient records, 688 participants). This analysis included all participants with at least 10 eligible cases during July 2012 – June 2015.

To ensure adequate statistical precision, the STS plans to report composite scores only for participants with at least 25 eligible cases during the 3-year measurement window. Thus, some of the analyses in this submission are limited to participants with at least 25 eligible cases.

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

***Note**: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. This quantity cannot be calculated directly because the "true" composite measure values are unknown, but may be estimated, as described below.

Calculation Details

Let θ_j denote the true unknown composite measure value for the j-th of J participants. Before estimating reliability, the numeric value of θ_j was estimated for each participant under the assumed hierarchical model. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. For each j, we randomly generated a large number (N) of possible numeric values of θ_j by sampling from the Bayesian posterior probability distribution of θ_j via MCMC sampling. Let $\theta_j^{(i)}$ denote the i-th of these N randomly sampled numerical values for the j-th participant.
2. For each j, the posterior mean $\hat{\theta}_j$ of θ_j was calculated as the arithmetic average of the randomly sampled values $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$; in other words $\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N} \theta_j^{(i)}$.

Our reliability measure was defined as the squared correlation between the set of hospital-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and the corresponding unknown true values $\theta_1, \ldots, \theta_J$. Let ρ^2 denote the unknown true squared correlation of interest and let $\hat{\rho}^2$ denote an estimate of this quantity. The estimate was calculated as

$$\hat{\rho}^2 = \frac{1}{N}\sum_{i=1}^{N} \rho_{(i)}^2$$

where

$$\rho_{(i)}^2 = \frac{\left[\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)\left(\hat{\theta}_j - \bar{\theta}\right)\right]^2}{\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)^2 \sum_{j=1}^{J}\left(\hat{\theta}_j - \bar{\theta}\right)^2}, \quad \bar{\theta} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N} \theta_j^{(i)} \quad \text{and} \quad \bar{\theta}^{(i)} = \frac{1}{J}\sum_{j=1}^{J} \theta_j^{(i)}.$$

A 95% Bayesian probability interval for ρ^2 was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers $\rho_{(1)}^2, \ldots, \rho_{(N)}^2$.

2a2.3. What were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

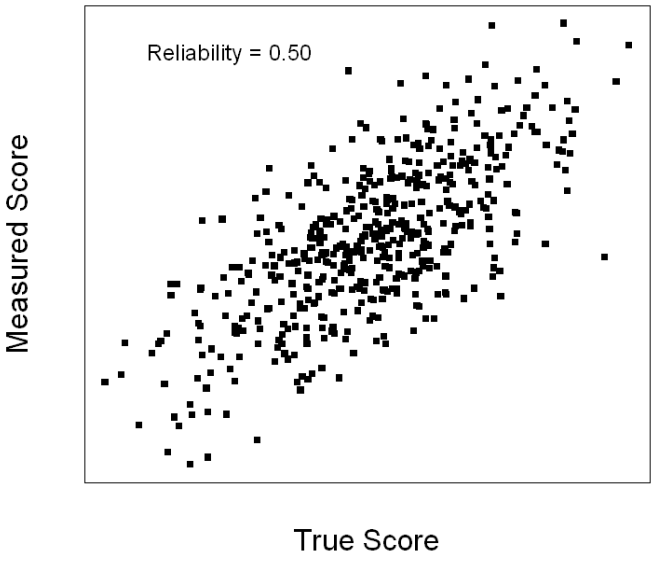
The estimated reliability of the STS MVRR+CABG composite measure using 3 years of data in participants with at least 25 total cases was 0.50 (95% CrI, 0.44 to 0.57), as outlined in the table below. For comparison, the reliability of the STS isolated CABG composite score was 0.77 (95% CrI, 0.74 to 0.80) using 1 year of data in 2013. Using 3 years of data from 2011 to 2013, the reliability of the STS AVR composite measure was 0.52 (95% CrI, 0.47 to 0.57), and the AVR+CABG measure was 0.50 (95% CrI, 0.45 to 0.54)

Time Span	Number of Participants Included	Number of Patients Included	Reliability $\hat{\rho}^2$(95% PrI)
3 years	703	24740	0.42 (0.35, 0.48)
3 years, participants with at least 25 cases	341	18924	0.50 (0.44, 0.57)
3 years, participants with at least 50 cases	143	12217	0.62 (0.52, 0.70)

Based in part on these results, we selected a threshold of 25 cases over 3 years, as a minimum threshold for receiving a site-specific STS MVRR+CABG composite score. This resulted in a reliability of 0.50 but reduced the number of programs eligible to receive a score from 703 to 341. A higher volume threshold would have yielded even higher reliability but at the cost of further reducing the number of programs eligible to receive a score.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

To interpret the results, we created a figure illustrating the accuracy of the measured scores when the true reliability = 0.50. Because the true score for the composite measure is unknown, we used simulated data with formula Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2,\ldots,341$ indicates the 341 participants and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure has a reliability of 0.50.



2b2. VALIDITY TESTING

***Note**: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include*

assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.

2b2.1. What level of validity testing was conducted?

☒ **Composite performance measure score**

☐ **Empirical validity testing**

☐ **Systematic assessment of face validity of <u>performance measure score</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

☐ **Systematic assessment of content validity**

☐ **Validity testing for component measures** *(check all that apply)*

***Note**: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Empirical validity testing of the component measure score(s)**

☐ **Systematic assessment of face validity of <u>component measure score(s)</u> as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

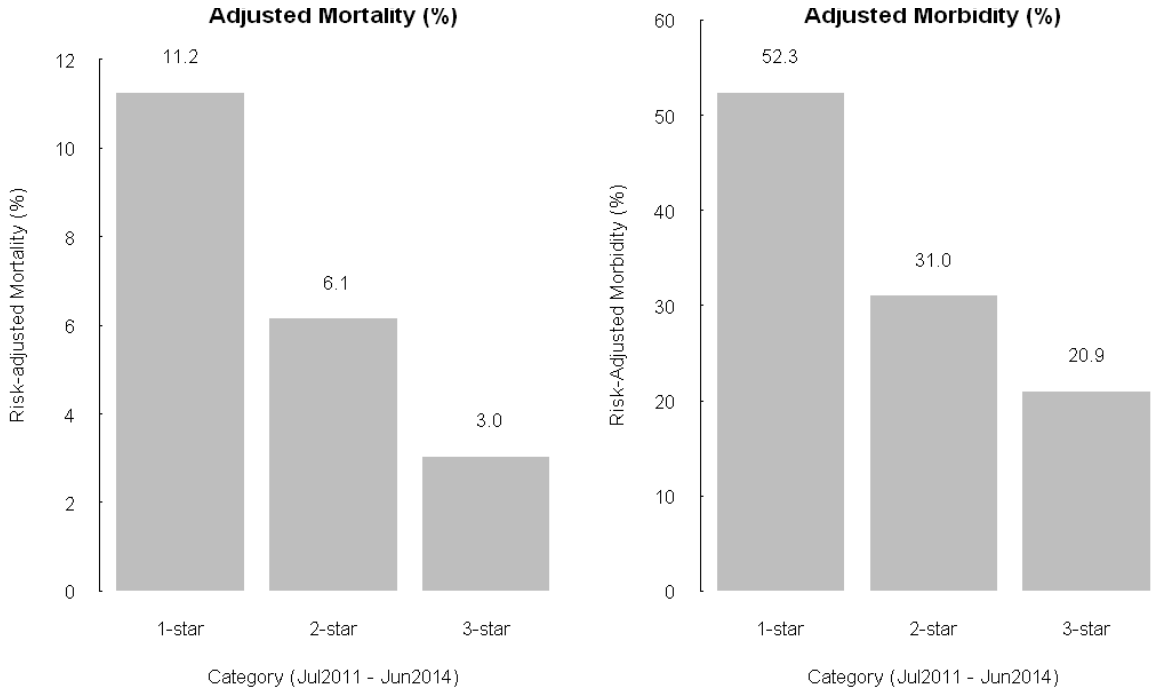
The tests on validity used the concept of performance categories to be more formally introduced in 2b5: Participants were labeled as having higher-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely above the overall STS average composite score. Participants were labeled as having lower-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely below the overall STS average composite score. Participants were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).

We compared risk-adjusted mortality and morbidity rates across the three performance groups. The measure has good face value if the three groups have different proportions as expected.

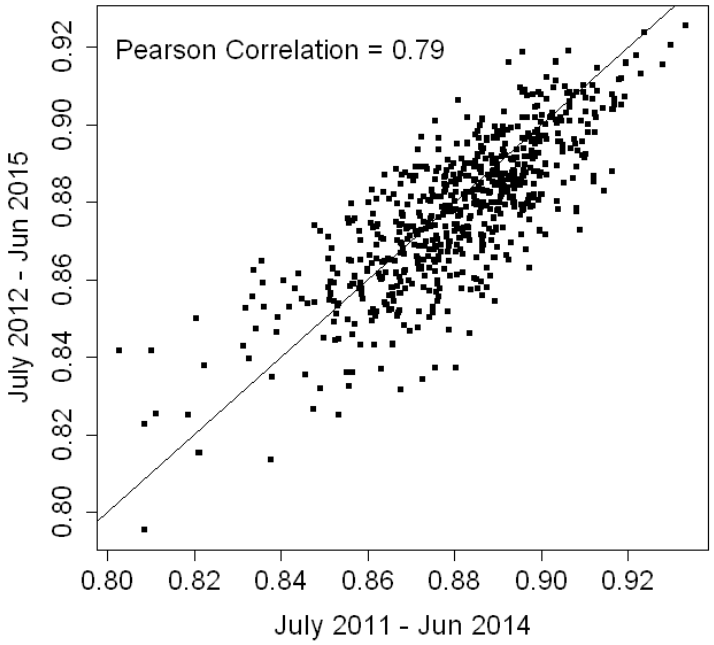
In addition, we assessed the extent to which a participant's composite score remains stable across two consecutive overlapping reporting periods. This analysis was restricted to 654 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (3.0% vs. 11.2%) and lower risk-adjusted morbidity (20.9% vs. 52.3%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS participants deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.

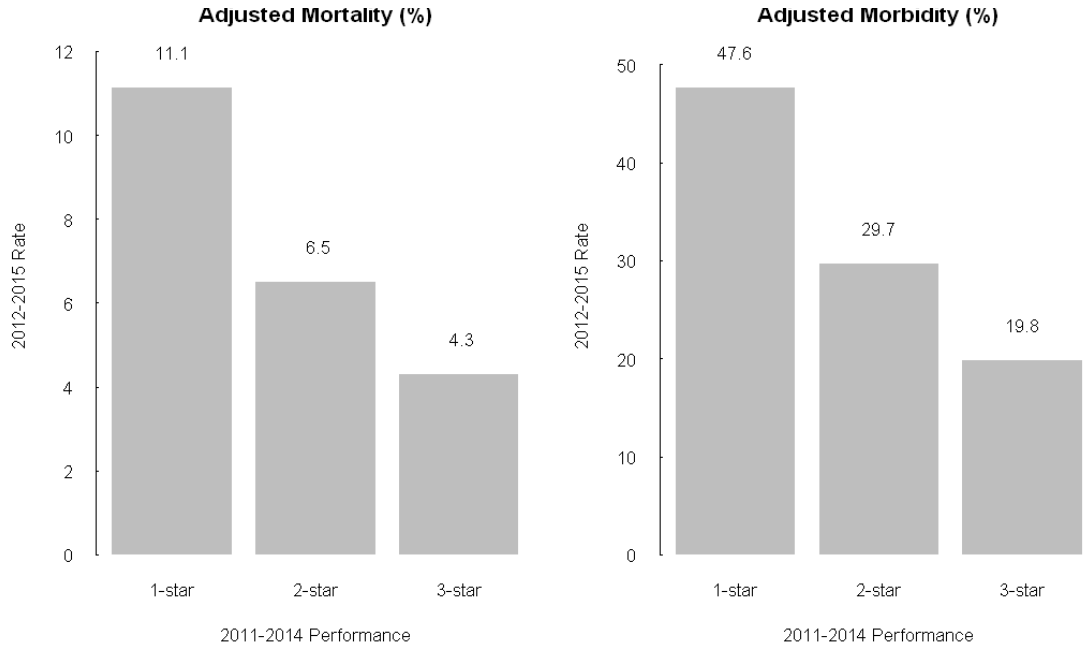


Stability of the composite measure over time was assessed in 654 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.



The Pearson correlation between the composite score calculated in the earlier and later time period was 0.79.

Using data from July 2012 – June 2015, we compared risk-adjusted mortality and morbidity rates across participants categories based on their composite measure performance in July 2011 – June 2014. Compared to 1-star participants, those with 3 stars had lower risk-adjusted mortality (4.3% versus 11.1%) and risk-adjusted morbidity (47.6% versus 19.8%) during July 2012 – June 2015.



2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.e., *what do the results mean and what are the norms for the test conducted?*)

The test results show that the composite measure behaves as expected and that results are reasonably consistent across two consecutive overlapping time periods. These results support the validity of the composite measure as a quality measure for MVRR + CABG procedures.

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

N/A

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

N/A

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☒ **Endorsed (or submitted) as individual performance measures**
☐ **No risk adjustment or stratification**
☒ **Statistical risk model**
☐ **Stratification by risk categories**
☐ **Other,** 14T

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

To adjust for case mix in the STS MVRR + CABG Composite Score [1], the published 2008 STS valve +CABG model [2] was modified and re-estimated in the current study population. The main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population who did not meet inclusion/exclusion criteria for the existing 2008 STS valve + CABG model. In addition, although the existing STS models predict the endpoints of "operative mortality" and "operative mortality or major morbidity" there is no existing model for predicting "major morbidity" as defined in the current study. In the future, as STS risk models are revised over time, the composite measure will be calculated with the most up to date STS risk model for the MVRR + CABG population.

Except where noted below, covariates for the modified operative mortality model were identical to the STS 2008 operative mortality model and covariates for the new major morbidity model were identical to the STS 2008 operative mortality or major morbidity model. For each of the two models, the list of covariates was modified as follows.

- Adjust for concomitant tricuspid repair. The STS 2008 models excluded patients undergoing a concomitant tricuspid procedure. Because the current study included patients undergoing concomitant tricuspid repair, an indicator variable for tricuspid repair was included.
- Adjust for tricuspid insufficiency using categories none or mild, moderate, and severe. The 2008 models included indicators of at least moderate tricuspid insufficiency. Because of the inclusion of operations with concurrent TV repair procedures, the surgeon panel felt it was necessary to more finely adjust the degrees of tricuspid insufficiency. The modified models include separate indicator variables for moderate tricuspid insufficiency and severe tricuspid insufficiency.
- Adjust for infectious endocarditis using categories active, treated, and none. The 2008 models include an indicator for active infections endocarditis but not for treated infectious endocarditis. The modified models include separate indicator variables for treated infectious endocarditis and active infectious endocarditis.

Considerations for adjusting for tricuspid repair

It is a generally accepted principle not to use what may be discretionary procedural decisions (e.g., whether or not to add a tricuspid valve repair) in profiling models. However, as discussed in the main article, there is accumulating evidence of the potential longitudinal merits of concomitant TVr. and the surgeon panel wanted to avoid discouraging the performance of this procedure by failing to account for its increased inherent risk of morbidity. Furthermore, the panel felt that the need to perform TVr may be a proxy for more advanced disease that may not be captured perfectly in the current STS data collection form.

References

1. Rankin JS, Badhwar V, He X, Jacobs JP, Gammie JS, Furnary AP, Fazzalari FL, Han J, O'Brien SM, Shahian DM. The Society of Thoracic Surgeons Mitral Valve Repair/Replacement plus Coronary Artery Bypass Grafting Composite Score: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force – This manuscript is currently being prepared for submission to The Annals of Thoracic Surgery.
2. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009 Jul;88(1 Suppl):S43-62.

2b4.4. What were the statistical results of the analyses used to select risk factors?

Estimated odds ratios from the modified STS 2008 models are summarized in the table below.

	Morbidity		Mortality	
Effect	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
Effects that do not interact with MV repair/replacements				
Preoperative atrial fibrillation	1.09 (1.02, 1.17)	0.0125	1.04 (0.92, 1.18)	0.4926
Race (v. others)				
Black	1.21 (1.08, 1.35)	0.0007	NA	.
Hispanic	1.16 (1.00, 1.35)	0.0529	NA	.
CVD (v. no)				
CVD with CVA	1.21 (1.10, 1.33)	0.0001	1.01 (0.86, 1.19)	0.9223
CVD without CVA	1.09 (0.98, 1.21)	0.1264	NA	.
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.16 (1.11, 1.21)	<.0001	1.16 (1.08, 1.26)	<.0001
Pre-op IABP or inotrope	2.21 (1.98, 2.47)	<.0001	1.43 (1.22, 1.69)	<.0001
Hypertension	1.11 (1.02, 1.20)	0.0189	NA	.
Immunosuppressive treatment	1.17 (1.02, 1.34)	0.0264	1.29 (1.02, 1.63)	0.0303
Peripheral vascular disease	1.08 (1.00, 1.17)	0.0536	1.28 (1.11, 1.48)	0.0007
MI (v. no recent MI)				
1-21 days	1.32 (1.23, 1.42)	<.0001	1.30 (1.13, 1.50)	0.0002
<=24 hrs	1.48 (1.16, 1.89)	0.0015	1.76 (1.28, 2.40)	0.0004
Number of previous operations (v. 0)				
1 previous operation	1.45 (1.15, 1.83)	0.0017	2.79 (1.88, 4.14)	<.0001
2 or more previous operations	1.50 (1.00, 2.24)	0.0485	2.68 (1.41, 5.06)	0.0025
Diabetes (v. no)				
Non-insulin diabetes	1.22 (1.12, 1.32)	<.0001	1.35 (1.17, 1.57)	<.0001
Insulin diabetes	1.08 (1.01, 1.16)	0.0233	1.10 (0.97, 1.24)	0.1565
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.10 (1.07, 1.14)	<.0001	1.16 (1.10, 1.22)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	2.17 (1.88, 2.50)	<.0001	2.66 (2.19, 3.23)	<.0001
Creatinine per 1 unit increase	1.62 (1.51, 1.73)	<.0001	1.46 (1.33, 1.61)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.20 (1.11, 1.29)	<.0001	1.39 (1.21, 1.59)	<.0001
Status (v. elective)				
Urgent	1.26 (1.18, 1.36)	<.0001	1.09 (0.96, 1.24)	0.1821
Emergent - no resuscitation	2.53 (1.75, 3.65)	<.0001	1.74 (1.12, 2.73)	0.0148
Emergent+resuscitation/Emergent Salvage	1.90 (1.07, 3.38)	0.0292	5.13 (2.83, 9.31)	<.0001
Active infections endocarditis	1.48 (1.20, 1.83)	0.0003	1.63 (1.18, 2.24)	0.0027
Treated infections endocarditis	0.91 (0.72, 1.16)	0.4538	0.57 (0.33, 0.97)	0.0393
Body surface area, m^2				
1.6 v. 2.0 in male	1.16 (1.01, 1.34)	0.0354	1.32 (1.02, 1.72)	0.0354
1.8 v. 2.0 in male	1.02 (0.97, 1.08)	0.4400	1.07 (0.97, 1.17)	0.1703
2.2 v. 2.0 in male	1.09 (1.05, 1.14)	<.0001	1.08 (1.01, 1.16)	0.0234
1.6 v. 1.8 in female	1.12 (1.06, 1.18)	0.0002	1.24 (1.12, 1.36)	<.0001
2.0 v. 1.8 in female	1.06 (1.00, 1.12)	0.0360	1.03 (0.94, 1.12)	0.5595
2.2 v. 1.8 in female	1.33 (1.15, 1.54)	0.0002	1.34 (1.06, 1.68)	0.0133

Time trend (half year increase)	0.98 (0.96, 1.00)	0.0541	1.03 (1.00, 1.06)	0.0440
Left main disease	NA	.	1.09 (0.96, 1.24)	0.1778
Unstable angina (no MI < 8days)	NA	.	1.01 (0.87, 1.17)	0.9382
Mitral stenosis	NA	.	1.21 (1.01, 1.46)	0.0399
Mitral insufficiency (>= moderate)	0.95 (0.86, 1.05)	0.3396	NA	.
Moderate tricuspid insufficiency (v. no-mild)	1.10 (1.02, 1.20)	0.0189	1.10 (0.96, 1.26)	0.1618
Severe tricuspid insufficiency (v. no-mild)	1.12 (0.98, 1.29)	0.1051	1.12 (0.89, 1.41)	0.3448
Mitral valve repair (v. replacement)	0.69 (0.59, 0.81)	<.0001	0.81 (0.59, 1.10)	0.1784
Tricuspid valve repair (v. none)	1.33 (1.19, 1.49)	<.0001	1.04 (0.85, 1.27)	0.7010
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs				
In MV replacements + CABG				
Age				
60 v. 50 (no reoperations, non-emergent)	1.16 (1.09, 1.23)	<.0001	1.70 (1.51, 1.91)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.35 (1.20, 1.52)	<.0001	2.88 (2.28, 3.64)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.57 (1.34, 1.84)	<.0001	4.84 (3.62, 6.49)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.15 (1.04, 1.28)	0.0063	1.14 (0.94, 1.37)	0.1794
CHF NYHA IV	1.36 (1.18, 1.55)	<.0001	1.49 (1.21, 1.83)	0.0002
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.04 (0.96, 1.14)	0.3436
Shock	2.07 (1.59, 2.69)	<.0001	1.89 (1.49, 2.39)	<.0001
In MV repairs + CABG				
Age				
60 v. 50 (no reoperations, non-emergent)	1.16 (1.10, 1.21)	<.0001	1.45 (1.31, 1.61)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.34 (1.21, 1.47)	<.0001	2.11 (1.72, 2.60)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.55 (1.36, 1.76)	<.0001	3.04 (2.35, 3.92)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.15 (1.05, 1.27)	0.0027	1.27 (1.06, 1.51)	0.0087
CHF NYHA IV	1.32 (1.18, 1.49)	<.0001	1.40 (1.14, 1.73)	0.0016
Shock	1.97 (1.56, 2.47)	<.0001	1.89 (1.49, 2.39)	<.0001
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.13 (1.06, 1.21)	0.0002

CHF = congestive heart failure; CVA = cerebrovascular accident (stroke); CVD = cardiovascular disease; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; NA = variable not used in model and estimate not available; NYHA = New York Heart Association.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

The modified models were assessed using data from 26,355 patients undergoing MVRR + CABG during July 2011 – June 2014.

Discrimination

To gauge discrimination, we calculated the c-statistics of both models. Bootstrapping was used to estimate and adjust for the "optimism" from estimating and evaluating the model on the same sample [1].

Calibration

The model fit was evaluated using 5-fold cross validation. The entire sample was randomly split into five equal sized groups. The calibration plot was created by following these steps:

1. One of the five groups was used as the testing sample
2. The other four groups were combined into the training sample
3. The revised model was estimated using the training sample
4. The expected probability of experience the event in the testing sample was calculated using the model estimated in step 3.

5. The expected probability (from step 4) and observed event rates were then compared in the testing sample and the calibration plot was created.

The above five steps were repeated five times so that each group was used as the testing sample once. In the end, we had five calibration plots for each model.

Reference

1. Harrell, F. E., Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine 15 (1996): 361-387.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

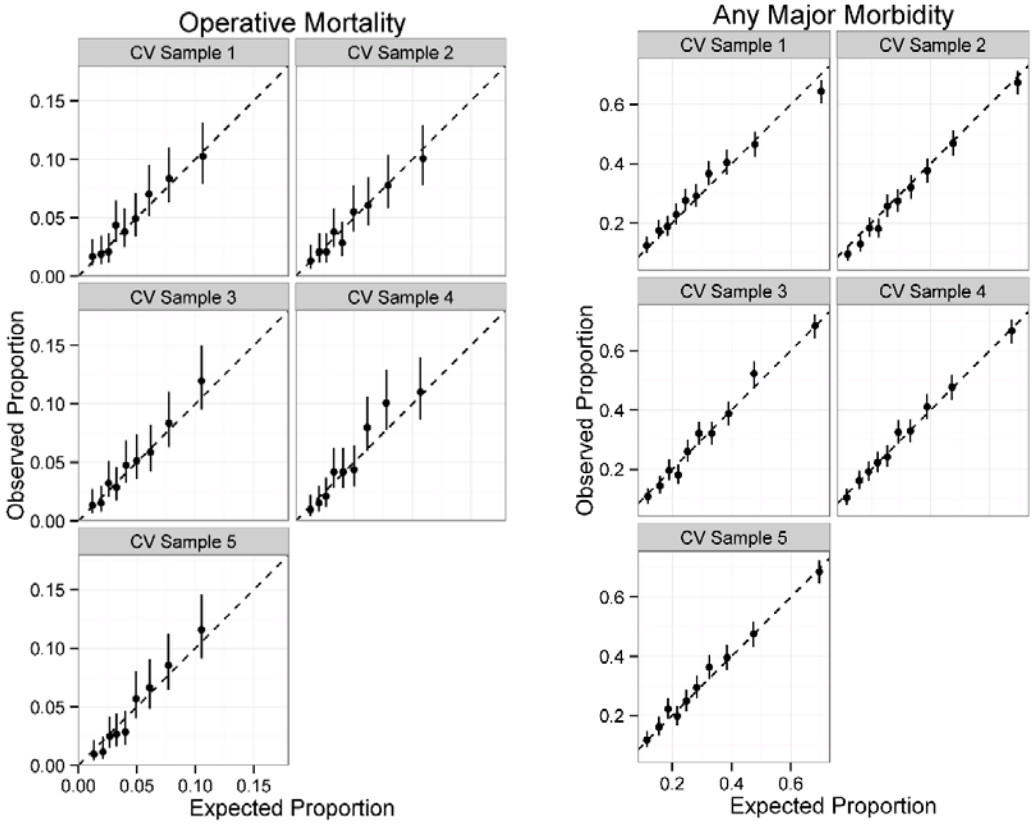
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The bootstrap-adjusted C statistic was 0.708 for the morbidity model and 0.738 for the mortality model. These numbers were comparable to the STS 2008 models when evaluated using the same sample (0.707 and 0.738 for morbidity and mortality endpoints, respectively.)

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A. The Hosmer-Lemeshow statistic was not calculated.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:



Plots of observed versus expected in cross validation samples, operative mortality

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted?*)

The results demonstrated that the STS cardiac surgery risk models are well calibrated and have good discrimination power. They are suitable for controlling differences in case-mix between centers.

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

Note: *Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The degree of uncertainty surrounding an STS participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS participant are reported along with a comparison to various benchmarks based on the national sample. Benchmarks include the overall average STS composite score and several percentiles (minimum, 10th, 25th, 75th, 90th, maximum). In addition, the composite measure result is converted into categories labeled as 1 to 3 stars. An STS participant receives 2 stars if the Bayesian credible interval surrounding their composite score overlaps the overall STS average. This rating implies that the STS participant's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the participant receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the participant receives 1 star (lower-than-expected performance).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Among participants with at least 25 cases over 3 years, around 91% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.

Performance categories

July 2011 – June 2014

	All Participants	**Participants N≥ 25**
Category	**Number of Participants, %**	**Number of Participants, %**
1-star	14, 2.0%	8, 2.3%
2-star	666, 94.7%	310, 90.9%
3-star	23, 3.3%	23, 6.7%

July 2012 – June 2015

	All Participants	**Participants N≥**

		25
Category	**Number of Participants, %**	**Number of Participants, %**
1-star	10, 1.5%	10, 2.9%
2-star	657, 95.5%	314, 91.0%
3-star	21, 3.1%	21, 6.1%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The Bayesian methodology allows direct probability interpretation of the results. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.*

If only one set of specifications for each component, this section can be skipped.

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted?*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

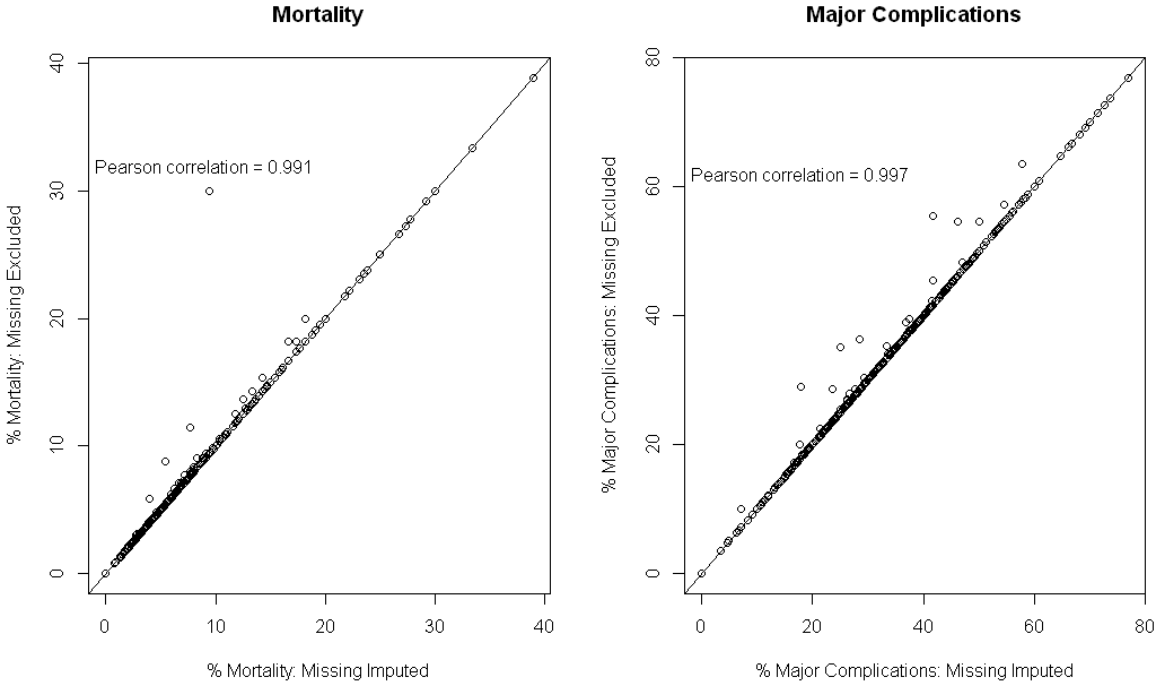
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.55% of records for operative mortality and 0.44% of

records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The overall frequency of missing data was 0.55% for operative mortality and 0.44% for major complications. The median participant-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications. The percent of participants with >10% missing data was 0.7% for mortality and 1.5% for major complications. As a sensitivity analysis, we re-calculated participant-specific mortality and complication rates after excluding records with missing data from the denominator. As shown in the figure below, there was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.



2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

These results suggest that our handling of missing outcome data is unlikely to impact performance results for the vast majority of participants.

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To verify that each domain contributes statistical information but does not dominate the composite, we calculated the correlations between each domain-specific estimate and the overall composite score. These analyses were performed using data from July 2011 – June 2014.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

Pearson Correlation With Overall Composite	
Mortality	**Morbidity**
0.60	0.91

The Pearson correlations were 0.60 for mortality versus the overall composite measure and 0.91 for morbidity domain score versus overall score.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

Although risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate. Both domains also contribute statistical information.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To form the composite, we rescaled the morbidity and mortality domains by dividing by their respective standard deviations across STS participants and then added the two domains together. This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains. To facilitate the assessment, we calculated for a 1 percentage point change in mortality, what percentage point change in morbidity would be needed to achieve the same impact on the composite measure.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.74 and 0.26, respectively. An implication of this weighting is that a 1 percentage point change in a participant's risk-adjusted mortality rate has the same impact as a 2.8 percentage point change in the site's risk-adjusted morbidity rate.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.e., *what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Data Collection:
There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor

- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
This is a new composite measure, which was developed in 2015 and will be published in 2016. STS plans to distribute participant-specific composite results in 2016 and roll out public reporting within the next year or so.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
Please see 4a.2.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Data are provided in 1b.2 and 1b.4 as required.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2014, 10% of participants were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0119 : Risk-Adjusted Operative Mortality for CABG
0120 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)
0121 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement
0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
1501 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** STS_MVRR_-_CABG_Composite_Score_Appendix_-_S.4-S.11-S.14-S.15-_1b.2-_1b.4-_manuscript.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
- Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 06, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** S.1. Measure-specific Web Page (Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)
Data collection forms: http://www.sts.org/sites/default/files/documents/STSAdultCVDataCollectionForm2_73_Annotated.pdf, http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;
Data specifications:
http://www.sts.org/sites/default/files/documents/word/STSAdultCVDataSpecificationsV2_73%20with%20correction.pdf, http://www.sts.org/sites/default/files/documents/STSAdultCVDataSpecificationsV2_81.pdf

Additional details are provided in the manuscript, which can be shared with NQF after it is published:
Rankin JS, Badhwar V, He X, Jacobs JP, Gammie JS, Furnary AP, Fazzalari FL, Han J, O'Brien SM, Shahian DM. The Society of Thoracic Surgeons Mitral Valve Repair/Replacement plus Coronary Artery Bypass Grafting Composite Score: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force.

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3017
De.2. **Measure Title:** PBM-02: Preoperative Hemoglobin Level
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure is designed to allow transfusion/blood use review committees to identify patients undergoing elective surgery with suboptimal, uncorrected hemoglobin levels that may have led to perioperative transfusion. This measure assesses, via stratification, pre-operative hemoglobin levels of selected elective surgical patients age 18 and over who received a perioperative red blood cell transfusion.
1b.1. **Developer Rationale:** There are many corrective interventions available for patients identified with preoperative sub-optimal hemoglobin levels in order to avoid a transfusion during or after the surgical procedure. As an essential component of blood management, pre-operative investigation and correction of anemia should be undertaken, since transfusion has been shown to increase adverse outcomes. Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need5.

One study of hip and knee arthroplasty patients found that those with a hemoglobin level <13.0g/dL. had four times the risk for blood transfusion than those with higher hemoglobin levels5.

Prevalence of preoperative anemia varies by population: Community-dwelling, >65 years old - <10%
i. Frail nursing home resident - >48%
ii. Surgical population – 5% to 75%
iii. Octogenarian, elective cardiac surgery – 49.4%1
iv. 7% of 9,462 patients undergoing total hip or total knee replacement2
v. >65 years old – 11% women, 10.2% men (NHANES Study)3
vi. Elective orthopedic surgery – 35%4

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

Preoperative anemia is also a predictor of postoperative transfusion in orthopedic, major colon, and major cardiac surgery. Since blood transfusion is the most frequently-performed hospital procedure (11% of hospital stays) and has increased by 126% from 1997 – 2010, and since blood transfusion can have adverse outcomes, such as prolonged length of stay and decreased functional status at discharge, investigation and correction of preoperative anemia is essential to any blood management program.

The World Health Organization has defined the levels of anemia for men at a hemoglobin measurement of less than 13.0, and for non-pregnant women at a hemoglobin measurement of less than 12.0. There has, however, been controversy over these levels. While there is debate regarding the hemoglobin level at which patients are considered anemic7, use of the WHO definition of anemia allows identification of patients for whom pre-operative investigation and correction of hemoglobin levels is warranted.

The intent of the measure is to provide information to providers and review groups about the incidence of transfusions in the various

strata, with the objective of identifying trends related to over- and underutilization of blood transfusions and correction of preoperative anemia.

5. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
6. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. The Journal of Bone and Joint Surgery. Volume 84-A – Number2 – February 2002.
7. Beutler E, Waalen J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? Blood Mar 1 2006 (107)5: 1747-1750.

S.4. **Numerator Statement:** Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
12.0 grams or above
>=11.0 and <12.0 grams (mild anemia)
>=8.0 and <11.0 grams (moderate anemia)
Below 8.0 grams (severe anemia)

S.7. **Denominator Statement:** Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.

S.10. **Denominator Exclusions:**
- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
- Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure.
- Patients with sickle cell disease or hereditary hemoglobinopathy

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

The developer provides the following path to support the relationship between the process of care (optimized hemoglobin levels prior to elective surgery) and outcomes:

1. Process: Optimized preoperative hemoglobin level

2. Elective surgical procedure performed
3. Reduced rate of blood transfusion
4. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

- The developer provided a guideline from the Network for Advancement of Transfusion Alternatives:
  - *Recommendation 2*: *We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl^{-1}, male ≥13 g dl^{-1}), according to the WHO criteria (Grade 2C). This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anaemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.* Grade 2C: Weak recommendation ("we suggest") and low or very low quality evidence (observational studies, randomized controlled tried with major limitations).
- The developer provided an additional 15 citations as sources of evidence for this measure, many of which were surgery type specific. Conclusions in the Fowler AJ et al. article published in 2015 state: "*Preoperative anaemia is associated with poor outcomes after surgery, although heterogeneity between studies was significant. It remains unclear whether anaemia is an independent risk factor for poor outcome or simply a marker of underlying chronic disease. However, red cell transfusion is much more frequent amongst anaemic patients*."

Guidance from the Evidence Algorithm

Based on SR/grading of clinical practice guideline (Box 3) → QQC provided (Box 4) → Moderate quality evidence based on additional relevant review articles provided (Box 5b) → MODERATE

Questions for the Committee:

- *Is the evidence directly applicable to the determination of preoperative hemoglobin level and the reduced risk of transfusion-related adverse outcomes?*
- *How strong is the evidence for this relationship?*

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Performance Gap

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Although there is no performance data on the measure as specified, the developer listed data with citations from the literature that indicates opportunity for improvement that relate to the focus of measurement.

Disparities

- The developed indicated that no disparity data are available.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support the Measure Focus

- Decreasing blood use is worthy
  This measure will show a large gap, but there is noise with the signal. patients with renal failure, for instance.
  Seems redundant with previous measure -- would be OK with one or the other, but not both
  Measure title not on list? What is 3017? Text is for Preop HGB testing; 3017 title is blood group testing

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
  - 12.0 grams or above
  - >=11.0 and <12.0 grams (mild anemia)
  - >=8.0 and <11.0 grams (moderate anemia)
  - Below 8.0 grams (severe anemia)
- Denominator Statement: Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.
- Denominator Exclusions:
  - Patients under age 18
  - Patients whose surgical procedure is performed to address a traumatic injury
  - Patients who have a solid organ transplant
  - Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
  - Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure
  - Patients with sickle cell disease or hereditary hemoglobinopathy
- Level of Analysis: Facility
- Care Setting: Hospital/Acute Care Facility
- No risk adjustment or risk stratification

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM;

Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously; **Bonnie results submitted**
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 78 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications, including the hemoglobin level thresholds, are consistent with evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.38
Denominator clearly describes the activity being measured	4.46

Numerator inclusions clear and appropriate	4.51
Denominator inclusions clear and appropriate	4.53
Numerator exclusions clear and appropriate	4.44
Denominator exclusions clear and appropriate	4.45
Accurately assesses the process of care to which it is addressed	4.13

This measure is being considered for trial use, thus full validity testing results are not expected and the Committee will not vote on this criterion.

2b3-2b7. Threats to Validity

2b3. Exclusions:
When data are available, the developer plans to analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Maternal and Fetal procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- ECMO procedures recorded in SNOMEDCT or ICD10PCS that start prior to the elective surgical procedure
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing any of the following conditions:
  - Traumatic Injury
  - Pregnancy, Childbirth, and the Puerperium
- Sickle Cell Disease and Related Blood disorders

Questions for the Committee:
- *Are there other threats to validity the measure developer should consider?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time

Question for the Committee:
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data
The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms None identified

Feedback :

None identified

***Questions for the Committee**:*

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 42T

Measure Title: PBM-02: Preoperative Hemoglobin Level

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 42T

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.

- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: 42T

☐Patient-reported outcome (PRO): 42T

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): 42T

☒ Process: Optimized hemoglobin levels prior to elective surgery.

☐ Structure: 42T

☐ Other: 42T

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

5. Process: Optimized preoperative hemoglobin level
6. Elective surgical procedure performed
7. Reduced rate of blood transfusion
8. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. *Br. Journ. Anesthesia*, 106 (1): 13-22 (2011).

http://bja.oxfordjournals.org/content/106/1/13.full

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 2: We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl^{-1}, male ≥13 g dl^{-1}), according to the WHO criteria (Grade 2C).

This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anaemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade 2C

Grading system

Strength of recommendation: is risk/benefit clear?

- Yes ⇨ strong recommendation=Grade 1: 'we recommend'
- No ⇨ weak recommendation=Grade 2: 'we suggest'

Quality of evidence

- High-quality evidence=A (meta-analyses, randomized controlled trials)
- Moderate-quality evidence=B (randomized controlled trials with limitations, observational studies with large effects)
- Low- or very low-quality evidence=C (obervational studies, randomized controlled tried with major limitations)

Grade of recommendation=6 possible grades

- Grade 1A
- Grade 1B
- Grade 1C
- Grade 2A
- Grade 2B
- Grade 2C

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

See above

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as 1a.4.1

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Detection, evaluation, and management of preoperative anemia in elective orthopedic surgery.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Grade C – low-quality evidence (observational studies, randomized control trials with major limitations).

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.4.3

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1966 – January 2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

5 observational studies, 3 cohort studies, 1 meta-analysis, 1 systematic literature review.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not stated in citation; in review of studies appears that 3 are small cohort studies.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Unstated in citation

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

None

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, MEDLINE and other relevant sources including professional association websites, The Cochrane Library, the National Guideline Clearinghouse, and other sources was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

1. American Red Cross: "Preoperative assessment and efforts to reduce the RBC transfusion requirement in the perioperative period include the evaluation and treatment of anemia prior to surgery and the evaluation for discontinuation or replacement of anticoagulant and antiplatelet medications …for a sufficient time prior to surgery in consultation with the prescribing physician." A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.
2. Society for Blood Management: The panel further recommended that the patient's target hemoglobin before elective surgery should be within the normal range (normal female >12 g/dL, normal male >13 g/dL). Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. *Anesth Analg* 2005;101:1858-61, p. 1860
3. "Preoperative anaemia is associated with poor outcomes after surgery" Fowler AJ et al. Meta-analysis of the association between preoperative anaemia and mortality after surgery. *Br J Surg* 2015 Oct;102(11):1314-24.

METHODS:

A systematic review and meta-analysis of observational studies exploring associations between preoperative anaemia and postoperative outcomes was performed. Studies investigating trauma, burns, transplant, paediatric and obstetric populations were excluded. The primary outcome was 30-day or in-hospital mortality. Secondary outcomes were acute kidney injury, stroke and myocardial infarction. Predefined analyses were performed for the cardiac and non-cardiac surgery subgroups. A post hoc analysis was undertaken to evaluate the relationship between anaemia and infection. Data are presented as odds ratios (ORs) with 95 per cent c.i.

RESULTS:

From 8973 records, 24 eligible studies including 949 445 patients were identified. Some 371 594 patients (39·1 per cent) were anaemic. Anaemia was associated with increased mortality (OR 2·90, 2·30 to 3·68; I(2) = 97 per cent; P < 0·001), acute kidney injury (OR 3·75, 2·95 to 4·76; I(2) = 60 per cent; P < 0·001) and infection (OR 1·93, 1·17 to 3·18; I(2) = 99 per cent; P = 0·01). Among cardiac surgical patients, anaemia was associated with stroke (OR 1·28, 1·06 to 1·55; I(2) = 0 per cent; P = 0·009) but not myocardial infarction (OR 1·11, 0·68 to 1·82; I(2) = 13 per cent; P = 0·67). Anaemia was associated with an increased incidence of red cell transfusion (OR 5·04, 4·12 to 6·17; I(2) = 96 per cent; P < 0·001). Similar findings were observed in the cardiac and non-cardiac subgroups.

CONCLUSION:

Preoperative anaemia is associated with poor outcomes after surgery, although heterogeneity between studies was significant. It remains unclear whether anaemia is an independent risk factor for poor outcome or simply a

marker of underlying chronic disease. However, red cell transfusion is much more frequent amongst anaemic patients.

4. British Committee for Standards in Haemotology: Recommendation: "Healthcare pathways should be structured to ensure anaemia screening and correction before surgery." Kotze A, Harris A, Baker C, Iqbal T, eta l. British Committee for Standards in Haemotology Guidelines on the Identification and Management of Pre-Operative Anemia. *British Journal of Haemotology* Volume 171, Issue 3: November 2015 pages 322-331.

5. Society for the Advancement of Blood Management: "Patients who are having a procedure for which preoperative screening is required are identified at least three to four weeks prior to surgery to allow sufficient time to diagnose and manage anemia, unless the surgery is of an urgent nature and must be performed sooner." (Standard 6.2) SABM Administrative and Clinical Standards for Patient Blood Management Programs, Third Edition. Unpublished work, 2014. Downloaded from www.SABM.org on April 9, 2016.

6. New York State Department of Health: "Careful evaluation of pre-existing anemia and its treatment prior to surgery are an effective strategy for reducing surgical transfusion requirements." New York State Council on Human Blood and Transfusion Services. Guidelines for Transfusion Options and Alternatives, 2010. Downloaded from www.wadswoth.org/labcert/blood_tissue July 2015.

7. 13 references (12 articles, one literature review) document increased rate of perioperative blood transfusion when preoperative anemia is present. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". *Ann Thorac Surg* 2007;83: 527 – 86.

8. 1 study of 296 elective orthopedic surgeries indicated through multivariate analysis that a significant relationship existed only between the need for transfusion and the preoperative hemoglobin level (p+ 0.00001) after hip and knee replacement. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. *The Journal of Bone and Joint Surgery*. Volume 84-A – Number2 – February 2002.

9. 1 systematic literature review of 29 included citations demonstrated that low hemoglobin and patient age were consistent risk factors for blood transfusion in orthopedic surgery. Barr PJ et al. Drivers of Transfusion Decision Making and Quality of the Evidence in Orthopedic Surgery: A Systematic Review of the Literature. *Transfusion Medicine Reviews*, Vol 25 No. 4 (October), 2011 pp. 304 – 316.
10. In a cohort study of 239 patients scheduled for transcatheter aortic valve implantation (TAVI), 62.3% were found to be anemic pre-procedurally and were referred to a blood conservation clinic (BCC) where they received a regimen of IV iron, oral iron, or epoetin alfa. Rates of transfusion in this cohort of 60 patients were assessed and compared with transfusion rates for TAVI patients prior to the initiation of the program. Implementation of the BCC was associated with a substantial decrease in the average blood transfusion rate from 33.3% before program initiation to 15.3% after implementation ($P < 0.001$). After adjusting for baseline hemoglobin values and comorbidities, being assessed at the BCC was strongly associated with a reduction in the need for transfusion (odds ratio, 0.28; 95% confidence interval, 0.11-0.69; P ¼ 0.006. Shuvy M, et al. Preprocedure Anemia Management Decreases Transfusion Rates in Patients Undergoing Transcatheter Aortic Valve Implantation. *Canadian Journal of Cardiology* (2016) Article in press.

11. A placebo-controlled, double-blind trial enrolling 316 patients scheduled for major, elective orthopedic hip or knee surgery who were expected to require 2.2 units of blood and who were not able or willing to participate in an autologous blood donation program examined the efficacy of Epogen treatment in reducing use of perioperative blood transfusion. Based on previous studies which demonstrated that pretreatment hemoglobin is a predictor of risk of receiving transfusion, patients were stratified into one of three groups based

on their pretreatment hemoglobin [-< 10 (n = 2) > 10 to 5 13 (n = 96), and > 13 to I 15 g/dL (n = 218)] and then randomly assigned to receive 300 Units/kg EPOGENQ 100 Units/kg EPOGEN@ or placebo by SC injection for 10 days before surgery, on the day of surgery, and for 4 days after surgery. All patients received oral iron and a low-dose post-operative warfarin. Treatment with EPOGENB 300 Units/kg significantly (p = 0.024) reduced the risk of allogeneic transfusion in patients with a pretreatment hemoglobin of > 10 to _< 13 g/dL; 5/31 (16%) of EPOGENB 300 Units/kg, 6126 (23%) of EPOGEN@ 100 Units/kg, and 13/29 (45%) of placebo treated patients were transfused. There was no significant difference in the number of patients transfused between EPOGENB (9% 300 Units/kg, 6% 100 Units/kg) and placebo (13%) in the > ***13*** to I ***15*** g/dL hemoglobin stratum. There were too few patients in the I 10 g/dL group to determine if EPOGEN@ is useful in this hemoglobin strata. In the > 10 to I 13 g/dL pretreatment stratum, the mean number of units transfused per EPOGENQ-treated patient (0.45 units blood for 300 Units/kg, 0.42 units blood for 100 Units/kg) was less than the mean transfused per placebo-treated patient (1.14 units) (overall p = 0.028). In addition, mean hemoglobin, hematocrit and reticulocyte counts increased significantly during the pre-surgery period in patients treated with EPOGEN. deAndrade JH, Jove M. Baseline Hemoglobin as a Predictor of Risk of Transfusion and Response to Epoetin alfa in Orthopedic Surgical Patients. *Am J of Orthoped*. 1996;25(8): 533-542.

12. Among 569 patients who underwent colorectal cancer surgery between 1998 and 2003, 32 anemic patients who received iron supplementation for at least 2 weeks preoperatively (group A) and 84 anemic patients who did not (group B) were studied.

There were no significant differences between groups A and B in age, sex, surgical technique, tumor stage, and operating time. Their Hgb and Hct values were similar at first presentation, but significantly different immediately before surgery (both P < 0.0001). There were no significant differences in intraoperative blood loss between the groups, but significantly fewer patients in group A needed an intraoperative blood transfusion (9.4% vs 27.4%, P < 0.05). Okuyama M et al. Preoperative iron supplementation and intraoperative transfusion during colorectal cancer surgery. *Surg Today*. 2005;35(1):36-40.

13. 1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that

a. The prevalence of preoperative anemia was 21-56%

b. Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.

Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. *Anesthesiology*, v. 113 No 2 August 2010.

14. A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Composite postoperative morbidity at 30 days was also higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

15. A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of women and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative

mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with a preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_02_evidence_attachment-635996215345848997.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
There are many corrective interventions available for patients identified with preoperative sub-optimal hemoglobin levels in order to avoid a transfusion during or after the surgical procedure. As an essential component of blood management, pre-operative investigation and correction of anemia should be undertaken, since transfusion has been shown to increase adverse outcomes. Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need5.

One study of hip and knee arthroplasty patients found that those with a hemoglobin level <13.0g/dL. had four times the risk for blood transfusion than those with higher hemoglobin levels5.

Prevalence of preoperative anemia varies by population: Community-dwelling, >65 years old - <10%

i. Frail nursing home resident - >48%
ii. Surgical population – 5% to 75%
iii. Octogenarian, elective cardiac surgery – 49.4%1
iv. 7% of 9,462 patients undergoing total hip or total knee replacement2
v. >65 years old – 11% women, 10.2% men (NHANES Study)3
vi. Elective orthopedic surgery – 35%4

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

Preoperative anemia is also a predictor of postoperative transfusion in orthopedic, major colon, and major cardiac surgery. Since blood transfusion is the most frequently-performed hospital procedure (11% of hospital stays) and has increased by 126% from 1997 – 2010, and since blood transfusion can have adverse outcomes, such as prolonged length of stay and decreased functional status at discharge, investigation and correction of preoperative anemia is essential to any blood management program.

The World Health Organization has defined the levels of anemia for men at a hemoglobin measurement of less than 13.0, and for non-pregnant women at a hemoglobin measurement of less than 12.0. There has, however, been controversy over these levels. While there is debate regarding the hemoglobin level at which patients are considered anemic7, use of the WHO definition of anemia allows identification of patients for whom pre-operative investigation and correction of hemoglobin levels is warranted.

The intent of the measure is to provide information to providers and review groups about the incidence of transfusions in the various strata, with the objective of identifying trends related to over- and underutilization of blood transfusions and correction of preoperative anemia.

5. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
6. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. The Journal of Bone and Joint Surgery. Volume 84-A – Number2 – February 2002.
7. Beutler E, Waalen J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? Blood Mar 1 2006 (107)5: 1747-1750.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This is a new measure for which approval for trial use is requested.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Incidence of preoperative anemia –
b. Incidence of anemia increases with age but varies by subpopulation.
i. Community-dwelling, >65 years old - <10%
ii. Frail nursing home resident - >48%
iii. Surgical population – 5% to 75%
iv. Octogenarian, elective cardiac surgery – 49.4%1
v. 7% of 9,462 patients undergoing total hip or total knee replacement2
vi. >65 years old – 11% women, 10.2% men (NHANES Study)3
vii. Elective orthopedic surgery – 35%4

8. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
9. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
10. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
11. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.
A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Composite postoperative morbidity at 30 days was also higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.
A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of women and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with a preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.
In addition, in a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 118 of the 141 respondents (81%) indicated that there was a gap in practice; 6 were not sure, and 17 reported no gap. Of the 118, most indicated that pre-operative anemia screening was done 3 or 4 days in advance of the elective surgical procedure. Given that 3-4 days is an insufficient period of time to correct any anemia, a high incidence of patients undergoing elective surgery with uncorrected anemia is presumed. Unpublished data.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

No disparities are identified.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparities are identified in the literature.

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Blood transfusion is the most common procedure performed during hospitalization,1 but research shows 50 percent of red blood cell transfusions are found to be inappropriate.2

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Most Frequent Procedures Performed in U.S. Hospitals, 2010, Healthcare Cost and Utilization Project (HCUP). February 2013. Agency for Healthcare Research and Quality.
2. Shander et al. Appropriateness of Allogeneic Red Blood Cell Transfusion: The International Consensus Conference on Transfusion Outcomes. Transfusion Medicine Reviews, Vol 25, No 3 (July), 2011: pp 232-246.e53.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not a PRO-PM

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.jointcommission.org/measure_development_initiatives.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-02_PreopHemoglobinLevel.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PreopHemoglobinLevel_v4_3_Wed_Jun_08_15.16.14_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
12.0 grams or above
>=11.0 and <12.0 grams (mild anemia)
>=8.0 and <11.0 grams (moderate anemia)
Below 8.0 grams (severe anemia)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Pre-operative hemoglobin level is represented as a code from the following value set and associated QDM datatype:
"Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Inpatient encounters are represented by the valueset and associated QDM datatype:
"Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
Selected elective surgical procedures are represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"
Transfusion of whole blood or packed cells is represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
- Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure.
- Patients with sickle cell disease or hereditary hemoglobinopathy

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

Traumatic injury is represented by a code from the following value set and associated QDM datatype:
Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Solid organ transplant is represented by a code from the following value set and associated QDM datatype;
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"

Pregnancy, delivered and not delivered, is represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: Maternal and Fetal Procedures" using "Maternal and Fetal Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.51)
Or
Attribute: "Diagnosis: Pregnancy, Childbirth, and the Puerperium Grouping Value Set (2.16.840.1.113762.1.4.1029.50)

ECMO is represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)"

Sickle cell disease and hereditary hemoglobinopathy is represented by a code from the following value set and associated QDM datatype:
Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*

Stratification 1 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result >= 12.0 g)"

Stratification 2 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 11.0 g)
(result < 12.0 g)

Stratification 3 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 8.0 g)
(result < 11.0 g)

Stratification 4 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result < 8.0 g)"

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Count
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Score within a defined interval

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not a PRO-PM or survey

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM_02_testing_form_for_trial_use.docx,PBM02_CMS601v0_Bonnie_Export.xlsx

National Quality Forum

Measure Testing Form for Trial Approval Program

Measure Title: PBM-02: Preoperative Hemoglobin Level
Date of Submission: 5/31/2016
Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submission that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide preliminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the measure meets reliability and validity must be in this form

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*including questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

1. DATA/SAMPLE USED FOR PRELMINARY TESTING OF THIS MEASURE

It is important that the measure developer use a data set to conduct preliminary testing in order to evaluate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

The Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.

Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: 42T	☐ other:

1.4. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

78 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Bonnie testing

was also performed for each stratum specified in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions, as well as complex cases containing multiple values for hemoglobin results as well as multiple transfusions, to confirm the stratification logic performed as expected.

If the Bonnie testing tool was used to provide a sample data set, please refer to the guidance for Bonnie testing found at this link: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80307 **Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. Refer to this link for an example of formatting Bonnie results:** http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=81576

Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.

RELIABILITY AND VALIDITY ASSESSMENTS

Note: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) **Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program.**

Include descriptions of:

Inter-abstractor reliability, and data element reliability of all critical data elements

Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-02.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.38
Denominator clearly describes the activity being measured	4.46
Numerator inclusions clear and appropriate	4.51
Denominator inclusions clear and appropriate	4.53
Numerator exclusions clear and appropriate	4.44
Denominator exclusions clear and appropriate	4.45
Accurately assesses the process of care to which it is addressed	4.13

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Denominator test cases positively test to ensure that only patients who have a blood transfusion administered <=5 day(s) after the start of selected surgical procedures are included in the denominator. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patients who receive transfusion one minute after surgery, at exactly 5 days, and 6 days after surgery. Patients receiving transfusions one minute and five days after surgery were included in the denominator, while patients receiving transfusions at 6 days after surgery were not.

Numerator test cases positively test to ensure patients who have a hemoglobin result recorded <= 45 days(s) prior to the start of surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in which hemoglobin

results were recorded >45 days prior to surgery or after surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures or encounter diagnoses. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-01 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Maternal and Fetal procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- ECMO procedures recorded in SNOMEDCT or ICD10PCS that start prior to the elective surgical procedure
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing any of the following conditions:
    - Traumatic Injury
    - Pregnancy, Childbirth, and the Puerperium
    - Sickle Cell Disease and Related Blood disorders.

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PBM02_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

n/a

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Usability and Use Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)

There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public

domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

n/a

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new measure.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance

results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
This is a new measure for which approval for trial use is requested.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
n/a

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630---

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of public comment and testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List

Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health

Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care
Medicine, Division of Vascular, Thoracic, Transplant Anesthesia
Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.
Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center
Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America
Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic
Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic

Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center
Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 05, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:** Ad.6. Copyright Statement

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-02: Preoperative Hemoglobin Level

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements that were deemed to be more difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"
3. "Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
4. "Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)"
5. "Procedure, Performed: Maternal and Fetal Procedures" using "Maternal and Fetal Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.51)"
6. "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"

7. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
8. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
9. Attribute: "Diagnosis: Pregnancy, Childbirth, and the Puerperium" using "Pregnancy, Childbirth, and the Puerperium Grouping Value Set (2.16.840.1.113762.1.4.1029.50)"
10. Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

Initial Population and Denominator Data Elements

Data elements 1- “Encounter, Performed: Encounter Inpatient,” 3- “Procedure, Performed: Blood Transfusion Administration” and 6- “Procedure, Performed: Selected Elective Surgical Procedures” are used to define the initial population and denominator of this measure.

On the feasibility scorecard, hospitals rated these data elements 1 and 6 as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year timeframe outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Four out of five hospitals rated capture of data element 6 as feasible or highly feasible, represented as a score of 2 or 3 out of 3. Facilities rating the data element as a 2 cited variation in clinical workflow and adoption of new technology as reasons for the lower rating. One site rated current state feasibility as a 1, as it did not currently have an interface between the OR scheduling system where this information was captured and the certified EHR technology. This site had plans to transition to an interfaced OR module in 1-2 years. All site rated the future state as highly feasible.

Finally, four out of five sites rated data element 3 as highly feasible, represented as a score of 3 out of 3. One hospital rated current feasibility as a 1 as blood transfusion was documented on a paper record, but had plans to implement blood transfusion via the EHR within six months.

Numerator Data Element

Data element 2- “Laboratory Test, Performed: Hemoglobin blood serum plasma” is used to define the numerator for this measure. Specifically, cases with a hemoglobin result recorded within 45 days prior to the start time of an elective surgical procedure meet numerator criteria. This time frame differs slightly from PBM-01, which evaluates hemoglobin results captured 14-45 days prior to surgery. In this measure, hemoglobin results up to the day of surgery meet numerator criteria.

Hospitals reported that hemoglobin results are routinely captured as structured data prior to surgery. However, limited interoperability between hospitals and their community partners, such as clinics and lab centers, limits the availability of structured data for lab results from external laboratories. Hospitals noted that many external results are received via fax, or as an electronic document, rather than in a format that can be structured and encoded in the EHR.

Hospitals rated feasibility of capturing this data element as high in most circumstances, but noted the interoperability issues may currently impact the availability of data for some cases.

Denominator Exclusions Data Elements

Data elements 4, 5, 7, 8, 9, and 10 are used to represent denominator exclusions.

Feasibility for data elements 4-“Procedure, Performed: ECMO,” 5-“Procedure Performed, Maternal and Fetal Procedures,” and 7-“Procedure, Performed: Solid Organ Transplant,” was found to be comparable to 3-“Procedure, Performed: Selected Elective Surgical Procedures.” These data elements are found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform solid organ transplant, which would not use this data element.

Data elements 8- “Attribute, Diagnosis: Traumatic Injury,” 9- “Attribute: Diagnosis: Pregnancy, Childbirth, and the Puerperium,” and 10-“Attribute: “Diagnosis: Sickle Cell Disease and Related Blood Disorders” represent encounter diagnoses. All hospitals rated these data elements as highly feasible. Discussion around these data elements suggested that the functionality to support collection of these data elements are well established.

Conclusion

Hospitals completing the feasibility scorecard largely reported the data elements required to calculate this measure to be feasible or highly feasible in the current state, with the exception of the numerator data element representing hemoglobin results. Capture of hemoglobin results from external laboratories will require improvements in interoperability or workarounds to support data collection. Approval for Trial Use status will support The Joint Commission’s efforts to further test this measure.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0127
De.2. **Measure Title:** Preoperative Beta Blockade
Co.1.1. **Measure Steward:** The Society of Thoracic Surgeons
De.3. **Brief Description of Measure:** Percent of patients aged 18 years and older undergoing isolated CABG who received beta blockers within 24 hours preceding surgery.
1b.1. **Developer Rationale:** This process measure seeks to improve the quality of care for patients undergoing isolated CABG. The use of preoperative beta blockers in isolated CABG is strongly associated with a reduction in postoperative atrial fibrillation. Postoperative atrial fibrillation leads to increase resource utilization, increases the risk of stroke, and independently predicts a lower long-term survival for CABG patients.

S.4. **Numerator Statement:** Number of patients undergoing isolated CABG who received beta blockers within 24 hours preceding surgery
S.7. **Denominator Statement:** Patients undergoing isolated CABG
S.10. **Denominator Exclusions:** Cases are removed from the denominator if preoperative beta blocker was contraindicated or if the clinical status of the patient was emergent or emergent salvage prior to entering the operating room.

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data : Registry
S.26. **Level of Analysis:** Clinician : Group/Practice, Facility

IF Endorsement Maintenance – Original Endorsement Date: May 09, 2007 **Most Recent Endorsement Date:** Jan 31, 2012

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?**

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary with **Summary of prior review reported in 2012**

- During previous review, the Committee agreed there was strong evidence to support the measure.
- The 2011 ACCF/AHA Guideline for Coronary Artery Bypass Graft Surgery includes the recommendations that "Beta blockers should be administered for at least 24 hours before CABG to all patients without contraindications to reduce the incidence or clinical sequelae of postoperative Atrial Fibrillation" *(Class I, Level of Evidence B)* and "Preoperative use of beta blockers in patients without contraindications, particularly in those with an LVEF greater than 30%, can be effective in reducing the risk of in-hospital mortality. *(Class IIa, Level of Evidence B)*
- Evidence submitted at this and the last review included meta analyses identifying some 30 randomized trials, observational studies, and a meta-analysis of 10 cardiac surgery trials that demonstrated an 82% reduction of post op VT/VF with use of beta blockers.

Changes to evidence from last review:
The developer provided updated evidence for this measure:

- Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.

The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review.

Questions for the Committee:

- *Does the Committee believe there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm: Process measure/systematic review and grading evidence (Box 3) → Information on QQC presented (Box 4) → SR conclusion (5b) → Moderate

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- At last review, the Committee identified the measure as important and, with a performance compliance mean of 84.8% (median, 86.6%), representative of a gap for which continued performance improvement was desirable.
- Most recent data is drawn from the 12 month period from October 2014 to September 2015. All eligible operations are included except those for which pre-op beta blockade was contraindicated and/or operative status was emergent/emergent salvage.
- Detail is provided regarding age, sex, race, ethnicity, insurance, bsa, acuity and more than 20 clinical conditions.
- Performance ranged from 0% to 100% with mean performance during this period at 93.5% (1,041 STS database participants; 134,689 operations).

Disparities

- The developer reports that for analysis of disparities, eligible patients from STS database participants with procedures between October 2011 and September 2015 were used.
- Relevant subgroups were defined by age, gender, race, ethnicity and insurance status.
- The performance ranges from 2011-2012 to 2014-2015 are presented below.
  Gender – Male, 93.01% - 94.73%; Female, 93.92% - 95.24%
  Age Groups - <75, 93.42% - 94.96%; >=75, 92.53% - 94.43%
  Race – White, 93.25% - 94.86%; Black, 94.75% - 95.98%; Other, 91.62%- 94.13%

Insurance - Age>=65, lowest performance 92.40% (Medicare+Medicaid) in earliest time period to 94.94% (Medicare w/o Medicaid/Commercial as highest in 2013 - 2014 period and for those in the Age<65, the low in earliest time period to high in most recent time period was 93.55% (Commercial/HMO) and 96.74% (None/Self Paid).

- The data suggests relatively uniform high use of preoperative beta blocker across all groups.

Questions for the Committee:

- *How does the Committee view the performance gap in the context pre-op beta blocker use in the population represented in the gap and the improvement since the measure was last endorsed?*
- *How should the disparities information be factored into consideration of sociodemographic factors going forward?*
- *Is the Committee aware of other disparities information that should be considered?*

Preliminary rating for opportunity for improvement: ☐ **High** **X** **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- Process measure, but with strong linkage to outcome.
- 88% last time, 93% now.
- Can anesthesiologists report this measure?
- Embarrassing that harmonization has not occurred
- This is a process measure, the percentage of patients receiving pre-operative beta-blockers prior to surgery. Evidence supportive for reducing the risk of post-op atrial fibrillation (Class I, level of evidence B) and risk of in-hospital mortality (Class IIa, level of evidence B).
- This indicates a moderate level of evidence. I don't think that this is new evidence, rather continued supportive evidence that doesn't require discussion or re-voting.
- This is a maintenance measure. The developer provided guidelines from 2011, which were the same guidelines used in the 2012 NQF review and re-endorsement of this measure. From what I can tell, there has been no significant new evidence that would impact these guidelines.

1b.

- Measure is demonstrating very high rates of compliance, most recent rates 94.9% overall with little to no variability and without measureable gaps to close by race, gender and insurance product. Because of the currently high rates of compliance, would consider reserve status for this measure.
- In the 2012 review, data was provided by the developer that showed only 84.8% of participants were meeting metric criteria. This improved to 93.5% during the period of October 2014 to September 2015. This is excellent improvement, but also still represents a performance gap.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The measure is specified for analysis at the group/practice and facility levels of analysis and is intended for use in the hospital/acute care setting.
- The measure assesses the number of patients (18 or older) undergoing isolated CABG who received beta blockers within 24 hours preceding surgery. Denominator exclusions are beta blocker contraindication or clinical status of the patient was emergent or emergent salvage prior to entering the operating room.
- The data source for the measure is the STS Adult Cardiac Surgery Database. Data is collected using the STS database collection form (version 2.81) that includes detailed items regarding a wide range of factors including procedure and

preoperative medications including whether beta blocker was prescribed either within 24 hours, not prescribed, or contraindicated.

- The measure is not risk adjusted or risk stratified.
- The apparent change since last endorsement is the denominator exclusion of emergent or emergent salvage.

Questions for the Committee :

- *Does the Committee have any question about the denominator exclusion related to emergent surgery effect on the measure performance, including when compared to performance at last endorsement when it was not listed as an exclusion?*
- *Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Previous reliability testing included inter-rater reliability testing of 40 randomly selected sites participating in the STS Adult Cardiac Surgery Database.

SUMMARY OF TESTING

Reliability testing level X Measure score ☐ Data element ☐ Both
Reliability testing performed with the data source and level of analysis indicated for this measure X Yes ☐ No

Method and Results of reliability testing

- Testing was done using the sample of 1,041 STS participants (134,689 operations) who submitted data between October 2014 and September 2015.
- Score level testing was done using a method described as an equivalent to signal-to-noise analysis. The developer notes that reliability increases with number of patients and the vast majority of STS participants have >30 eligible patients per year thus calculating reliability with 30 patients per participant provides a conservative lower bound for the actual reliability that will be achieved when the measure is applied to STS data from a 1 year period.
- The developer states that the estimated reliability with 30 eligible patients per participant = 0.78. The table below provides information about the minimum sample sizes to achieve for reliability at 3 additional levels along with the percent of STS database participants who meet the minimum sample size.

	Reliability 0.50	Reliability 0.60	Reliability 0.70
Minimum required sample size per participant	8	13	20
Percent of participants meeting minimum sample size	99%	98%	97%

- The developer interprets the results as adequate statistical reliability for use in confidential feedback and public reporting.

Questions for the Committee:

- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm Precise specifications (Box 1) → Empiric reliability testing (Box 2) →Testing at measure score level (Box 4) → Method described and appropriate (Box 5) → Level of confidence (Box 6)

Preliminary rating for reliability: ☐ High X Moderate ☐ Low ☐ Insufficient

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. X Yes ☐ Somewhat ☐ No

Question for the Committee:

○ *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

- At prior maintenance submission, validity testing was conducted using the STS database audit process that involved 40 randomly selected sites (of the 628 eligible database participants that had at least 100 eligible cases for the measure and reported data for each month during January 1, 2009 – December 31, 2009).

SUMMARY OF TESTING

Validity testing level ☐ Measure score ☐ Data element testing against a gold standard X Both

Method of validity testing of the measure score:

☐ Face validity only
X Empirical validity testing of the measure score

Validity testing method:

- Data element testing was done using the STS Adult Cardiac Surgery Database Audit. Ten percent of database participants (107) were audited. Auditing involved re-abstraction of data for 20 cases from each audited participant and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates were calculated for each variable, each variable category, and overall. Overall aggregate agreement rate was 96.17% .
- For validity testing and comparison of participants overtime, STS participants with procedures during both October 2013 – September 2014 and October 2014 – September 2015 were used.
- Performance measure score testing was done using face validity and predictive validity - assessing stability of performance over time. ***There is some disagreement about whether stability in performance demonstrates predictive validity; some would argue that changes in performance over time are to be expected—and are, in fact, desirable—as the result of quality improvement interventions. NQF guidance suggests that predictive validity should compare measure results to another measure of the same construct or to a different outcome measure.***
- Participants were placed into one of three groups based on 95% exact binomial confidence interval (CI) of its observed proportion – low performance (95% exact binomial CI of event rate entirely below population average), high performance (95% CI entirely above 1) and mid performance (remaining participants). Predictive validity analysis was restricted to 1,015 participants that received the measure in both time periods (10/2013 – 9/2014 and 10/2014 – 9/2015). To assess impact of the beta blocker contraindication, the distribution of the measure with and without the exclusion was computed.
- **Results:** Of high performers in the earlier period, 77% were high performers in the second period while only 12% of mid performers in the early period became high performers in the second period; i.e., participants that performed better than average in the earlier period were over 6 times more likely to be better performers in the next year. Also, it is noted that movement from low to high was proportionately greater than the reverse. Aggregated proportion of patients receiving preoperative beta blocker in the later period (10/2014 – 9/2015) was also calculated and is reported as 84.3% (low performance), 95.1% (mid performance), and 98.8% high performance with the conclusion that the measure reflects the proportion of patients who received preoperative beta blockade and that the past measurement can be used to predict future performance.

Questions for the Committee:

- *Does the Committee agree that the analysis provided demonstrates validity of the measure?*
- *Do the results allow conclusions about quality that can be translated to meaningful interpretation of performance by users?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The exclusions to the measure are contraindication to use of beta blockers and emergent/emergent salvage.
- The developer has calculated the distribution of the participant-specific observed proportion of patients receiving beta blockers with and without the exclusion for the 10/2014 – 9/2015 time period and shown the percent change across each of the three performance groups to demonstrate its effect on measure results. With beta blocker contraindication and emergent/emergent salvage status exclusion criteria applied, proportion of patients receiving beta blockers changed as follows:
  - low performance participants with exclusion, 18.9%; without, 19.6%;
  - mid performance 51.7% with exclusion; without, 57.6%;
  - high performance 29.4%; without, 22.8%.
  - The developer reports that STS database participants performing better and worse than the STS average has remained similar over the two time periods, 10/2013 – 9/2014 and 10/2014 – 9/2015, with more than 50% having performance indistinguishable from the STS average.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** **X None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

As noted in earlier sections, the developer has presented information about low, mid, and high performance participants.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed. The measure uses a single data source and has one set of specifications.

2b7. Missing Data

Overall rate of missing data is reported as 0.1%. Missing data are imputed to "no" (preoperative beta blocker); participants with greater than 5% (7 of 1,048 participants) missing data are excluded from the measure calculation. The developer reported that 99% of participants had 4% or lower missing data.

Guidance from the Validity Algorithm Specifications consistent with evidence (Box 1) → Threats to validity (Box 2) → Empirical validity testing (Box 3) → Validity systematically assessed (Box 4) → Confidence that scores are indicator of quality (Box 5)

Preliminary rating for validity: ☐ **High** **X Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1.

- Data elements clearly defined as a part of the extensive STS registry. Exclusion for emergent surgery cases seems reasonable for a process measure that involves medication administration 24 hours preceding the procedure.
- The measure has been assessed in the past as having appropriate specificity. There are two points that I would like clarification on, however, 1) is a "contraindication" simply the presence of a statement, at the physician's discretion, that says a beta blocker is contraindicated or does it require that a defined set of accepted contraindicatory criteria are met? and 2) beta blocker dose is not specified. Is this also left to provider discretion? Are the benefits of giving beta blockers the same at all dosages or is there a narrow range of widely accepted dosing for beta blockers?

2a2.

- Reliability scores increase appropriately with increased sample size. Reliability testing was completed at the score level with an estimated score of 0.78 for 30 cases.
- Data provided by the developer shows the measure to be highly reliable.

2b1.

- No concerns with validity related to specifications.
- Just general question about the method for validity scoring … were the groups deemed as low, mid or high performers based on other measures or was it based on past performance of this measure (which started off being pretty high and is now high overall). It would make sense if performance was compared based on other measures in the CABG portfolio.
- During the last discussion, the use of predictive validity testing was called into question. Other methods of validity testing were also employed however and the measure does appear to have adequate validity.

2b2.

- No concerns with validity testing.
- Yes, the scope of validity testing is adequate: Ten percent of database participants (107) were audited. Auditing involved re-abstraction of data for 20 cases from each audited participant and comparison of 82 individual data elements with those submitted to the data warehouse.

2b3.

- Exclusions are consistent with the evidence.
- This measure currently has very high rates of performance and does not demonstrate meaningful differences, at least in the data provided. Going out to the STS public reporting website does demonstrate a composite ""Receipt of Required Perioperative Medications"" which does demonstrate variation between practices (star ratings), but it is unclear what is included in the composite or which medication measures could be impacting the rates.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.

- Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants. Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.

- There are no additional costs for data collection specific to the measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time

- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database

Participation Agreement. STS analyses indicate that the STS database includes more than 90% of cardiothoracic programs in the US.

Questions for the Committee:
Is the effort and cost associated with abstracting the required data elements appropriate to the value of the measure?

Preliminary rating for feasibility: ☐ **High** **X Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

- These data elements are generated as a routine part of care. For this particular process measure that is not risk adjusted, a practice not participating in the STS registry could feasibly collect data for this measure in a quality improvement effort.
- This is a simple and highly feasible metric to implement, document, and track.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? **X Yes** ☐ **No**

Current use in an accountability program? **X Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Unexpected findings (positive or negative) during implementation: None identified

Potential harms
- The developer reports it is not aware of any negative unintended consequences, noting that all public reporting has potential for such things as gaming and risk aversion. The developer attempts to control gaming though its audit process and risk aversion by accounting for the expected risk for providers who care for sicker patients.

Feedback:
- At the time of endorsement reported in the June 2012 Surgery Endorsement Maintenance report, public comment noted that the measure could be used as a composite with #0126 Selection of Antibiotic Prophylaxis for Cardiac Surgery Patients. The developer stated that the denominator of 0127 differed from that of 0126. The committee noted that endorsement as an individual measure does not preclude its use in a composite.

Questions for the Committee*:*
- *Does the measure continue to be useful for improvement given the high level of performance?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

- Measure is one of 11 measures in the CABG composite measure, publicly reported on the STS website. This

individual measure is used in the PQRS program. As a stand-alone process measure with an overall rate of 94.9%, it is hard to say that there is gap and opportunity for improvement.
- Measures are reported through the STS database which has a large-scale program aimed at improving the quality of cardiothoracic surgery nationwide.

Criterion 5: Related and Competing Measures
Harmonization and Related or competing measures • Measures 0117 and 0127 are STS measures of beta blocker use that are harmonized. CMS Measure #0284, Surgery Patients on Beta-Blocker Therapy Prior to Arrival Who Received a Beta-Blocker During the Perioperative Period, is a similar, though more broad, measure, based on administrative claims data for use at the facility or regional levels. The CMS measure has a greater number of exclusions and extensive data collection algorithm. Uses of the measures differ. Other related measures include the STS measures that comprise the STS CABG Composite Score.

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0127

Measure Title: Preoperative Beta Blockade

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: 0696 STS CABG Composite Score

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome:

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: preoperative beta blockade

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

Process – Preoperative Beta Blocker administration – Outcome – reduced post-operative atrial fibrillation – reduction in long-term mortality and reduced resource consumption.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and URL for guideline (*if available online*):

Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, et al. 2011 ACCF/AHA guideline for coronary artery bypass graft surgery. Circulation 2011;124:e652-735.

http://circ.ahajournals.org/content/124/23/e652

1a.4.2. Identify guideline recommendation number and/or page number and quote verbatim, the specific guideline recommendation.

Page e152

4.5. Perioperative Beta Blockers: Recommendations

Class I Recommendation

Beta blockers should be administered for at least 24 hours before CABG to all patients without contraindications to reduce the incidence or clinical sequelae of postoperative atrial fibrillation. (Level of Evidence: B)

Class IIa Recommendation

Preoperative use of beta blockers in patients without contraindications, particularly in those with an LVEF greater than 30%, can be effective in reducing the risk of in-hospital mortality.

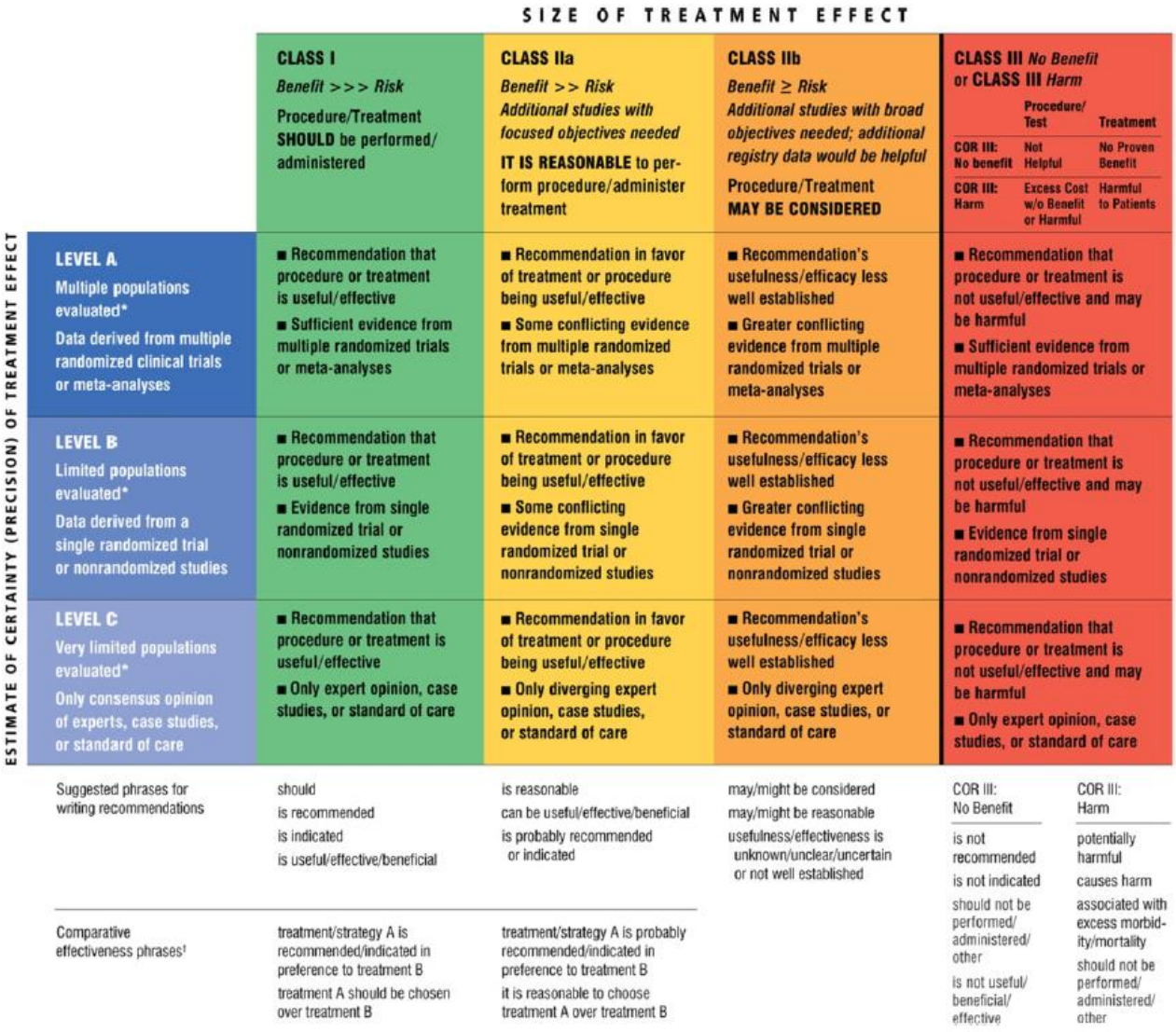
(Level of Evidence: B)

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Class 1 Level B. Recommendation that procedure or treatment is useful/effective. Evidence from single randomized trial or nonrandomized studies

Class IIa Level B. Recommendation in favor of treatment or procedure being useful/effective. Some conflicting evidence from single randomized trial or nonrandomized studies

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)



SIZE OF TREATMENT EFFECT

ESTIMATE OF CERTAINTY (PRECISION) OF TREATMENT EFFECT	CLASS I *Benefit >>> Risk* Procedure/Treatment **SHOULD** be performed/ administered	CLASS IIa *Benefit >> Risk* *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	CLASS IIb *Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	CLASS III *No Benefit* or CLASS III *Harm* Procedure/Test / Treatment COR III: No benefit — Not Helpful / No Proven Benefit COR III: Harm — Excess Cost w/o Benefit or Harmful / Harmful to Patients
LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical trials or meta-analyses	■ Recommendation that procedure or treatment is useful/effective ■ Sufficient evidence from multiple randomized trials or meta-analyses	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Sufficient evidence from multiple randomized trials or meta-analyses
LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is useful/effective ■ Evidence from single randomized trial or nonrandomized studies	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Evidence from single randomized trial or nonrandomized studies
LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	■ Recommendation that procedure or treatment is useful/effective ■ Only expert opinion, case studies, or standard of care	■ Recommendation in favor of treatment or procedure being useful/effective ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation's usefulness/efficacy less well established ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Only expert opinion, case studies, or standard of care
Suggested phrases for writing recommendations	should is recommended is indicated is useful/effective/beneficial	is reasonable can be useful/effective/beneficial is probably recommended or indicated	may/might be considered may/might be reasonable usefulness/effectiveness is unknown/unclear/uncertain or not well established	COR III: No Benefit: is not recommended; is not indicated; should not be performed/ administered/ other; is not useful/ beneficial/ effective COR III: Harm: potentially harmful; causes harm; associated with excess morbidity/mortality; should not be performed/ administered/ other
Comparative effectiveness phrases†	treatment/strategy A is recommended/indicated in preference to treatment B treatment A should be chosen over treatment B	treatment/strategy A is probably recommended/indicated in preference to treatment B it is reasonable to choose treatment A over treatment B		

A recommendation with Level of Evidence B or C does not imply that the recommendation is weak. Many important clinical questions addressed in the guidelines do not lend themselves to clinical trials. Although randomized trials are unavailable, there may be a very clear clinical consensus that a particular test or therapy is useful or effective.

*Data available from clinical trials or registries about the usefulness/efficacy in different subpopulations, such as sex, age, history of diabetes, history of prior myocardial infarction, history of heart failure, and prior aspirin use. †For comparative effectiveness recommendations (Class I and IIa; Level of Evidence A and B only), studies that support the use of comparator verbs should involve direct comparisons of the treatments or strategies being evaluated.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

ACCF/AHA Task Force on Practice Guidelines. Methodologies and Policies from the ACCF/AHA Task Force on Practice Guideline. June 2010.

http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf and

http://circ.ahajournals.org/site/manual/index.xhtml

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Information relevant to this section is provided in detail in the previous sections and the referenced guideline. Please see Appendix.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Preoperative beta blockers have been shown to reduce the incidence of postoperative atrial fibrillation in CABG patients. Observational analyses suggest that preoperative beta blocker use is associated with a reduction in perioperative deaths. Preoperative use of beta blockers in patients without contraindications, particularly in those with an LVEF greater than 30%, can be effective in reducing the risk of in-hospital mortality.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

See 1a.7

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.7

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

See 1a.7

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

See 1a.7

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

See 1a.7

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

See 1a.7

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

See 1a.7

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

See 1a.7

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
1a._Evidence_-_0127_Preoperative_Beta_Blockade.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This process measure seeks to improve the quality of care for patients undergoing isolated CABG. The use of preoperative beta blockers in isolated CABG is strongly associated with a reduction in postoperative atrial fibrillation. Postoperative atrial fibrillation leads to increase resource utilization, increases the risk of stroke, and independently predicts a lower long-term survival for CABG patients.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare
Affects large numbers, Frequently performed procedure, A leading cause of morbidity/mortality, High resource use, Severity of illness, Patient/societal consequences of poor quality
1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
CABG is a frequently performed procedure and a large number of patients undergo CABG yearly in the US. The development of post-operative atrial fibrillation consumes excess resources.

Beneficial pharmacological effects of beta blockers
Beta blockers have pleiotropic effects, many of which are likely to reduce the incidence of adverse cardiac events following cardiac surgery [1-4]. These agents reduce sympathetic nervous system activity; they are anti-arrhythmic, and they decrease heart rate, systolic blood pressure, and myocardial contractility. These effects will in turn reduce myocardial oxygen consumption and mitigate supply-demand mismatch, one cause of perioperative ischemia, infarct and death.

Beta blockers may reduce shear stress and stabilize vulnerable plaques, another mechanism by which they might reduce the likelihood of infarction, and they may increase the threshold for VF associated with ischemia [5]. Some have postulated that beta blockade may mitigate perioperative inflammatory processes and subsequent rapid progression of coronary plaque, which may explain why several studies have shown a long-term reduction in cardiac events with perioperative beta blockade [4;6;7] beyond the acute postoperative period.

Preoperative beta blockade in cardiac surgery
The most compelling justification for preoperative beta blockade use, and its inclusion as a performance measure for cardiac surgery, is its impact on the development of postoperative atrial fibrillation. This common complication occurs in about 22% of patients undergoing isolated CABG surgery by STS Database participants, and it results in increased resource utilization (LOS). The Virginia Cardiac Surgery Quality Initiative (VCSQI) found that atrial fibrillation added an average 10.3% ($2,744) and 2.2 days length of stay to a typical isolated CABG hospitalization [8]. Postoperative atrial fibrillation increases the risk of stroke [9-11], an often devastating complication, as well as other thromboembolic complications. It may produce hemodynamic compromise in some patients and at the very least is symptomatically unpleasant. It is a common cause of hospital readmission [12], and multiple studies show that the development of postoperative atrial fibrillation is an independent predictor of long-term survival following CABG surgery [13-17].

Meta-analyses have identified almost thirty randomized trials demonstrating a significant reduction in the incidence of atrial fibrillation following cardiac surgery, usually CABG [18-20]. This complication occurs much more frequently following heart surgery than non-cardiac surgery because of features such as pre-existing conduction system disease, sympathetic activation and increased endogenous catecholamines, cannulation, cardiac manipulations, pericardial inflammation, cardiac fluid shifts, cooling and rewarming of the heart, cardioversion, cardioplegia, cardiopulmonary bypass, and the use of inotropic agents. These marked differences from non-cardiac surgery probably explain why the incidence of atrial fibrillation is greater in cardiac surgery, and why non-cardiac patients do not appear to have a reduction in their already low incidence of this complication with beta blockade [19]. These factors are not eliminated even if adequate revascularization is achieved. Because of the substantial reduction in the incidence of atrial fibrillation in almost all cardiac surgery trials, use of these agents for this indication is a longstanding ACCF/AHA Class 1 (A) recommended therapy for patients without complications, and a similar recommendation has been published by the American College of Chest Physicians [21].

A second rationale for use of preoperative beta blockade in cardiac surgery was demonstrated by Ferguson and colleagues in a 2002

study [22]. This observational study included 629,877 patients in the STS Adult Cardiac Surgery Database between 1996 and 1999. Patients who received beta-blockers had decreased short-term mortality risk using both adjustment for patient risk and center effects (OR, 0.94; 95% CI, 0.91-0.97) and treatment propensity matching (OR, 0.97; 95% CI, 0.93-1.00). However, among patients with ejection fraction less than 30%, preoperative beta blockade was associated with a non-significant trend towards higher mortality (OR, 1.13; 95% CI, 0.96-1.33; P =.23). Interestingly, this study also showed a trend towards reduced stroke rate, which contrasts with findings previously noted for non-cardiac surgery. This is consistent with results from the study of Amory and colleagues [23] and may result from both the anti-arrhythmic effects of these drugs and direct neuroprotective effects. Finally, two smaller observational studies from Belgium and Australia have also demonstrated a reduction in CABG mortality with preoperative beta blockade [24;25]. For all these reasons, beta blockers may be useful to reduce mortality and ischemia in CABG patients with EF > 30%, but not patients with EF < 30%.

Finally, a recent meta-analysis of ten cardiac surgery trials demonstrated an 82% reduction of postoperative VT/VF with the use of beta blockers [19].

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Poldermans D, Devereaux PJ. The experts debate: perioperative beta-blockade for noncardiac surgery--proven safe or not? Cleve Clin J Med 2009 Nov;76 Suppl 4:S84-S92.
2. Schouten O, Bax JJ, Dunkelgrun M, Feringa HH, Poldermans D. Pro: Beta-blockers are indicated for patients at risk for cardiac complications undergoing noncardiac surgery. Anesth Analg 2007 Jan;104(1):8-10.
3. Mangano DT. Perioperative cardiac morbidity. Anesthesiology 1990 Jan;72(1):153-84.
4. Yeager MP, Fillinger MP, Hettleman BD, Hartman GS. Perioperative beta-blockade and late cardiac outcomes: a complementary hypothesis. J Cardiothorac Vasc Anesth 2005 Apr;19(2):237-41.
5. Poldermans D, Boersma E, Bax JJ, Thomson IR, van d, V, Blankensteijn JD, et al. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. N Engl J Med 1999 Dec 9;341(24):1789-94.
6. Mangano DT, Layug EL, Wallace A, Tateo I. Effect of atenolol on mortality and cardiovascular morbidity after noncardiac surgery. Multicenter Study of Perioperative Ischemia Research Group. N Engl J Med 1996 Dec 5;335(23):1713-20.
7. Poldermans D, Boersma E, Bax JJ, Thomson IR, Paelinck B, van de Ven LLM, et al. Bisoprolol reduces cardiac death and myocardial infarction in high-risk patients as long as 2 years after successful major vascular surgery. European Heart Journal 2001 Aug 1;22(15):1353-8.
8. Speir AM, Kasirajan V, Barnett SD, Fonner E Jr. Additive costs of postoperative complications for isolated coronary artery bypass grafting patients in Virginia. Ann Thorac Surg 2009 Jul;88(1):40-5.
9. D´Agostino RS, Svensson LG, Neumann DJ, Balkhy HH, Williamson WA, Shahian DM. Screening carotid ultrasonography and risk factors for stroke in coronary artery surgery patients. Ann Thorac Surg 1996 Dec;62(6):1714-23.
10. Likosky DS, Leavitt BJ, Marrin CA, Malenka DJ, Reeves AG, Weintraub RM, et al. Intra- and postoperative predictors of stroke after coronary artery bypass grafting. Ann Thorac Surg 2003 Aug;76(2):428-34.
11. Stamou SC, Hill PC, Dangas G, Pfister AJ, Boyce SW, Dullum MK, et al. Stroke after coronary artery bypass: incidence, predictors, and clinical outcome. Stroke 2001 Jul;32(7):1508-13.
12. D´Agostino RS, Jacobson J, Clarkson M, Svensson LG, Williamson C, Shahian DM. Readmission after cardiac operations: prevalence, patterns, and predisposing factors. J Thorac Cardiovasc Surg 1999 Nov;118(5):823-32.
13. Villareal RP, Hariharan R, Liu BC, Kar B, Lee VV, Elayda M, et al. Postoperative atrial fibrillation and mortality after coronary artery bypass surgery. J Am Coll Cardiol 2004 Mar 3;43(5):742-8.
14. Mariscalco G, Klersy C, Zanobini M, Banach M, Ferrarese S, Borsani P, et al. Atrial fibrillation after isolated coronary surgery affects late survival. Circulation 2008 Oct 14;118(16):1612-8.
15. El-Chami MF, Kilgo P, Thourani V, Lattouf OM, Delurgio DB, Guyton RA, et al. New-onset atrial fibrillation predicts long-term mortality after coronary artery bypass graft. J Am Coll Cardiol 2010 Mar 30;55(13):1370-6.
16. Filardo G, Hamilton C, Hebeler RF, Jr., Hamman B, Grayburn P. New-onset postoperative atrial fibrillation after isolated coronary artery bypass graft surgery and long-term survival. Circ Cardiovasc Qual Outcomes 2009 May;2(3):164-9.
17. Bramer S, van Straten AH, Soliman Hamad MA, Berreklouw E, Martens EJ, Maessen JG. The impact of new-onset postoperative atrial fibrillation on mortality after coronary artery bypass grafting. Ann Thorac Surg 2010 Aug;90(2):443-9.
18. Crystal E, Connolly SJ, Sleik K, Ginger TJ, Yusuf S. Interventions on prevention of postoperative atrial fibrillation in patients undergoing heart surgery: a meta-analysis. Circulation 2002 Jul 2;106(1):75-80.
19. Wiesbauer F, Schlager O, Domanovits H, Wildner B, Maurer G, Muellner M, et al. Perioperative beta-blockers for preventing surgery-related mortality and morbidity: a systematic review and meta-analysis. Anesth Analg 2007 Jan;104(1):27-41.
20. Burgess DC, Kilborn MJ, Keech AC. Interventions for prevention of post-operative atrial fibrillation and its complications after cardiac surgery: a meta-analysis. European Heart Journal 2006 Dec 1;27(23):2846-57.
21. Bradley D, Creswell LL, Hogue CW, Jr., Epstein AE, Prystowsky EN, Daoud EG. Pharmacologic prophylaxis: American College of

Chest Physicians guidelines for the prevention and management of postoperative atrial fibrillation after cardiac surgery. Chest 2005 Aug;128(2 Suppl):39S-47S.
22. Ferguson TB, Jr., Coombs LP, Peterson ED. Preoperative beta-blocker use and mortality and morbidity following CABG surgery in North America. JAMA 2002 May 1;287(17):2221-7.
23. Amory DW, Grigore A, Amory JK, Gerhardt MA, White WD, Smith PK, et al. Neuroprotection is associated with beta-adrenergic receptor antagonists during cardiac surgery: evidence from 2,575 patients. J Cardiothorac Vasc Anesth 2002 Jun;16(3):270-7.
24. ten Broecke PW, De Hert SG, Mertens E, Adriaensen HF. Effect of preoperative beta-blockade on perioperative mortality in coronary surgery. Br J Anaesth 2003 Jan;90(1):27-31.
25. Weightman WM, Gibbs NM, Sheminant MR, Whitford EG, Mahon BD, Newman MA. Drug therapy before coronary artery surgery: nitrates are independent predictors of mortality and beta-adrenergic blockers predict survival. Anesth Analg 1999 Feb;88(2):286-91.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular, Surgery, Surgery : Cardiac Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Medication Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;
http://www.sts.org/sites/default/files/documents/STSAdultCVDataSpecificationsV2_81.pdf

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
None

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of patients undergoing isolated CABG who received beta blockers within 24 hours preceding surgery

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator – 24 hours preceding surgery
Denominator – 12 months

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of isolated CABG procedures in which preoperative beta blockers [MedBeta (STS Adult Cardiac Surgery Database Version 2.81)] is marked "yes"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Patients undergoing isolated CABG

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Number of isolated CABG procedures excluding cases for which preoperative beta blockers were contraindicated or if the clinical status of the patient was emergent or emergent salvage prior to entering the operating room. The SQL code used to create the function used to identify cardiac procedures is provided in the Appendix.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Cases are removed from the denominator if preoperative beta blocker was contraindicated or if the clinical status of the patient was emergent or emergent salvage prior to entering the operating room.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Procedures with preoperative beta blockers [MedBeta (STS Adult Cardiac Surgery Database Version 2.81)] marked as "Contraindicated" or procedures with Status [Status(STS Adult Cardiac Surgery Database Version 2.81)] marked "Emergent" or "Emergent Salvage"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
N/A

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please refer to numerator and denominator sections for detailed information.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
The source fields required by the preoperative beta blockade measure had only 0.1% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with greater than 5% missing data are excluded from the calculation of the measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database Version 2.81

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
2.1_Testing_-_0127_Preoperative_Beta_Blockade.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0127
Measure Title: Preoperative Beta Blockade
Date of Submission: 6/5/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

STS Adult Cardiac Surgery Database (ACSD) Version 2.81

1.3. What are the dates of the data used in testing?

October 2014 – September 2015

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The calculation of the preoperative beta blockade measure of the 12 months from October 2014 to September 2015 used 134,689 operations from 1,041 STS participants.

Distribution of participant sample sizes (denominator), and observed proportion of patients receiving the measure (numerator/denominator)

Stat	N	% Pre-operative Beta Blockade
N	1041.0	1041.0
Mean	129.4	93.5
STD	103.6	9.3
IQR	110.0	8.2
0%	1.0	0.0
10%	36.0	81.6
20%	52.0	88.9
30%	67.0	92.9
40%	83.0	95.3
50%	99.0	97.1
60%	121.0	98.3
70%	148.0	99.2
80%	193.0	100.0
90%	263.0	100.0
100%	826.0	100.0

Distribution of participants by geographic regions

REGION	
Midwest	214
Northeast	104
South	321
West	146

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

All eligible isolated operations were included except cases with reasons for not receiving preoperative beta blockade: contraindication and/or operative status emergent/emergent salvage.

	Effects	Overall N=134689
Age (years)	Median (IQR)	66.0 (58.0, 72.0)
	Missing	0 (0.0%)
Sex	Male	101,106 (75.1%)
	Female	33,525 (24.9%)
	Missing	58 (0.0%)

	Effects	Overall N=134689
Race - Asian	No	127,643 (94.8%)
	Yes	4,173 (3.1%)
	Missing	2,873 (2.1%)
Race - Black / African American	No	121,502 (90.2%)
	Yes	10,316 (7.7%)
	Missing	2,871 (2.1%)
Race - White	No	20,280 (15.1%)
	Yes	111,598 (82.9%)
	Missing	2,811 (2.1%)
Race - American Indian / Alaskan Native	No	130,948 (97.2%)
	Yes	872 (0.6%)
	Missing	2,869 (2.1%)
Race - Other	No	127,146 (94.4%)
	Yes	4,283 (3.2%)
	Missing	3,260 (2.4%)
Native Hawaiian / Pacific Islander	No	131,136 (97.4%)
	Yes	623 (0.5%)
	Missing	2,930 (2.2%)
Hispanic or Latino Ethnicity	No	118,172 (87.7%)
	Yes	9,536 (7.1%)
	Missing	6,981 (5.2%)
Insurance: Younger than 65	Medicare/Medicaid	17,049 (27.7%)
	Commercial/HMO	36,519 (59.3%)
	None/Self Paid	4,779 (7.8%)
	Other	3,203 (5.2%)
Insurance: 65 or Older	Medicare+Medicaid	4,640 (6.3%)
	Medicare+Commercial without Medicaid	40,271 (55.1%)
	Medicare without Medicaid/Commercial	28,228 (38.6%)
Region	NORTHEAST	21,323 (15.8%)
	SOUTH	58,907 (43.7%)
	MIDWEST	33,129 (24.6%)
	WEST	21,330 (15.8%)
Body Surface Area (m)	<1.5	1,742 (1.3%)
	>=1.5 and <1.75	16,155 (12.0%)
	>=1.75 and <2	46,010 (34.2%)
	>=2	70,740 (52.5%)
	Missing	42 (0.0%)
Diabetes	No Diabetes	68,449 (50.8%)
	Diabetes - Noninsulin	40,187 (29.8%)
	Diabetes - Insulin	24,680 (18.3%)

	Effects	Overall N=134689
	Diabetes - Other	366 (0.3%)
	Diabetes - Missing Treatment	748 (0.6%)
	Missing	259 (0.2%)
Hypertension	No	14,182 (10.5%)
	Yes	120,269 (89.3%)
	Missing	238 (0.2%)
Renal Function	Creatinine <1 mg/dL	64,859 (48.2%)
	Creatinine 1-1.5 mg/dL	54,226 (40.3%)
	Creatinine 1.5-2 mg/dL	8,145 (6.0%)
	Creatinine 2-2.5 mg/dL	1,728 (1.3%)
	Creatinine >2.5 mg/dL	1,337 (1.0%)
	Dialysis	4,170 (3.1%)
	Missing	224 (0.2%)
Dyslipidemia	No	15,411 (11.4%)
	Yes	118,770 (88.2%)
	Missing	508 (0.4%)
Chronic Lung Disease (CLD)	None	96,629 (71.7%)
	Mild	14,736 (10.9%)
	Moderate	6,778 (5.0%)
	Severe	5,864 (4.4%)
	5	6,556 (4.9%)
	Missing	4,126 (3.1%)
Peripheral Vascular Disease (PVD)	No	114,549 (85.0%)
	Yes	19,318 (14.3%)
	Missing	822 (0.6%)
Cerebrovascular Disease (CVD)	No CVD	107,064 (79.5%)
	CVD-NO CVA	27,625 (20.5%)
Endocarditis	No Endocarditis	134,468 (99.8%)
	Treated Endocarditis	64 (0.0%)
	Active Endocarditis	8 (0.0%)
	Endocarditis - Missing Type	7 (0.0%)
	Missing	142 (0.1%)
Acuity Status	Elective	53,395 (39.6%)
	Urgent	81,273 (60.3%)
	Missing	21 (0.0%)
Myocardial Infarction	No Prior MI	63,990 (47.5%)
	MI >21 days	26,426 (19.6%)
	MI 8-21 days	6,866 (5.1%)
	MI 1-7 days	33,881 (25.2%)
	MI 6-24 hrs	1,833 (1.4%)
	MI <= 6 hrs	245 (0.2%)
	MI - Missing Timing	335 (0.2%)
	Missing	1,113 (0.8%)

	Effects	Overall N=134689
Cardiogenic Shock	No	134,087 (99.6%)
	Yes	559 (0.4%)
	Missing	43 (0.0%)
Preop IABP	No	127,879 (94.9%)
	Yes	6,644 (4.9%)
	Missing	166 (0.1%)
Congestive Heart Failure	No CHF	107,721 (80.0%)
	CHF NYHA-I	2,254 (1.7%)
	CHF NYHA-II	7,944 (5.9%)
	CHF NYHA-III	9,566 (7.1%)
	CHF NYHA-IV	5,170 (3.8%)
	CHF Missing NYHA	916 (0.7%)
	Missing	1,118 (0.8%)
Number of Diseased Coronary Vessels	None	114 (0.1%)
	One	5,340 (4.0%)
	Two	25,867 (19.2%)
	Three	102,438 (76.1%)
	Missing	930 (0.7%)
Left Main Disease > 50%	No	45,830 (34.0%)
	Yes	41,891 (31.1%)
	Missing	46,968 (34.9%)
Ejection Fraction (%)	Median (IQR)	55.0 (45.0, 60.0)
	Missing	3,697 (2.7%)
Aortic Stenosis	No	128,102 (95.1%)
	Yes	4,084 (3.0%)
	Missing	2,503 (1.9%)
Mitral Stenosis	No	131,430 (97.6%)
	Yes	704 (0.5%)
	Missing	2,555 (1.9%)
Tricuspid Stenosis	No	131,601 (97.7%)
	Yes	85 (0.1%)
	Missing	3,003 (2.2%)
Pulmonic Stenosis	No	130,419 (96.8%)
	Yes	33 (0.0%)
	Missing	4,237 (3.1%)
Aortic Insufficiency	None	87,571 (65.0%)
	Trivial	12,975 (9.6%)
	Mild	10,314 (7.7%)
	Moderate	2,037 (1.5%)

	Effects	Overall N=134689
	Severe	78 (0.1%)
	N/A or Not Documented	20,679 (15.4%)
	Missing	1,035 (0.8%)
Mitral Insufficiency	None	42,368 (31.5%)
	Trivial	32,978 (24.5%)
	Mild	31,929 (23.7%)
	Moderate	8,462 (6.3%)
	Severe	575 (0.4%)
	N/A or Not Documented	17,532 (13.0%)
	Missing	845 (0.6%)
Tricuspid Insufficiency	None	44,673 (33.2%)
	Trivial	38,625 (28.7%)
	Mild	25,029 (18.6%)
	Moderate	3,992 (3.0%)
	Severe	328 (0.2%)
	N/A or Not Documented	20,852 (15.5%)
	Missing	1,190 (0.9%)
Pulmonic Insufficiency	None	70,867 (52.6%)
	Trivial	20,232 (15.0%)
	Mild	6,472 (4.8%)
	Moderate	556 (0.4%)
	Severe	42 (0.0%)
	N/A or Not Documented	35,035 (26.0%)
	Missing	1,485 (1.1%)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

We used the same dataset of isolated CABG operations from October 2014 to September 2015 for the entire report. The three exceptions are:

1. For validity testing and the comparison of participants over time, we used STS participants with procedures during both October 2013 - September 2014 and October 2014 - September 2015 time periods.
2. For the analysis of population disparities, current and over time, we used eligible patients from STS participants with procedures between October 2011 and September 2015 and defined relevant subgroups by age, gender, race, ethnicity and insurance status.
3. For the analysis on the impact of exclusions, we included the cases with contraindication for preoperative beta blockade and operative status emergent/emergent salvage.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

We report trends of preoperative beta blockade among the following groups: Age (<75, ≥75), Gender, Race (White, Black and Other), Hispanic Ethnicity and Insurance (<65, ≥65).

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a measure that is due to true between-unit differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). Equivalently, it is the squared correlation between a measurement and the true value. For this NQF submission, the measurement of interest is each participant's observed proportion. The true value is the proportion that would be observed hypothetically if the sample size was very large (i.e. infinite).

For the j-th participant, let n_j denote the number of eligible patients, let $y_j$denote the number of patients receiving beta-blockers, and let $\bar{y}_j = y_j/n_j$ denote the proportion of patients receiving beta-blockers. In addition, let μ_j denote the underlying true value of $\bar{y}_j$. To estimate reliability, we assumed the following hierarchical model for the data. At the first stage of the hierarchy, we assume that y_j is distributed according to a binomial distribution with sample size n_j and probability parameter μ_j. At the second stage of the hierarchy, we assumed that μ_j varies across participants according to a Beta distribution with mean $E[\mu_j] = \alpha/(\alpha+\beta)$ and $\text{var}[\mu_j] = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$, where α and β are unknown parameters to be estimated from the data. The unknown parameters α and β were estimated via maximum likelihood using the BETABIN macro for SAS software (BETABIN, version 2.2, 2005. Qi Statistics). The sample for this analysis included all **1,041 participants** and **134,689 eligible patients** in the main study period October 2013-September 2014. After estimating α and β, we then calculated the reliability that would be achieved if the measure were to be calculated on a sample size of 30 patients per participant. This estimated reliability was calculated as

$$\text{reliability} = [\text{corr}(\bar{y}, \mu)]^2 = \frac{1}{1 + (\hat{\alpha} + \hat{\beta})/n}$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote maximum likelihood estimates of α and β, respectively, and n =30. Because reliability increases with n, and because the vast majority of STS participants have >30 eligible patients per year, the reliability calculated with n =30 patients per participant provides a conservative lower bound for the actual reliability that will be achieved when the measure is applied to STS data from a 1 year period. Using the above formula, we also calculated the sample size n required per participant to achieve reliability of at least 0.50, 0.60, and 0.70, and the proportion of STS participants with at least this number of eligible patients in the most recent 1-year testing sample.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Estimated parameter values of the beta distribution were $\hat{\alpha}$= 7.9296 and $\hat{\beta}$= 0.5301. The estimated reliability with 30 eligible patients per participant was 1/(1+ (7.9296 + 0.5301)/30)=0.78.

Based on these estimated parameter values, a sample size of 8 eligible patients per participant is needed to attain reliability of 0.50 and a sample size of 20 eligible patients per participant is needed to attain reliability of 0.70. During October 2014-September 2015, 99% of STS participants met the minimum required sample size for 0.50 reliability and 97% of STS participants met the minimum required sample size for 0.70 reliability.

	Reliability 0.50	Reliability 0.60	Reliability 0.70
Minimum required sample size per participant	8	13	20
Percent of participants meeting minimum sample size	99%	98%	97%

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Reliability is comparable to or better than other NQF-endorsed STS outcome measures. The proposed measure has adequate statistical reliability to be used for confidential feedback reporting as well as public reporting.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☒ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Critical data elements

Participating sites are randomly selected for participation in STS Adult Cardiac Surgery Database Audit, which is designed to evaluate the accuracy, consistency, and comprehensiveness of data collection and ultimately validate the integrity of the data contained in the database. Telligen has conducted audits on behalf of STS since 2006. In 2015, 10% of STS Adult Cardiac Surgery Database participants (N=107) were audited. The audit process involves re-abstraction of data for 20 cases and comparison of 82 individual data elements with those submitted to the data warehouse. Agreement rates are calculated for each of the 82 variables, each variable category and overall. In 2015 the overall aggregate agreement rate was 96.17%, demonstrating that the data contained in the STS Adult Cardiac Surgery Database are both comprehensive and highly accurate.

Performance measure score

We calculated and compared the observed proportions of patients receiving the measure in the three performance groups. The measure has good face value if the three groups have different proportions as expected.

Face validity also implies that the measure is regarded as useful and valid by its intended users, including providers, consumers, payers, and regulators. The measure was developed with a panel of surgeon experts and statisticians. We have had near-universal acceptance of this composite by all stakeholders, with few if any relevant suggestions for change.

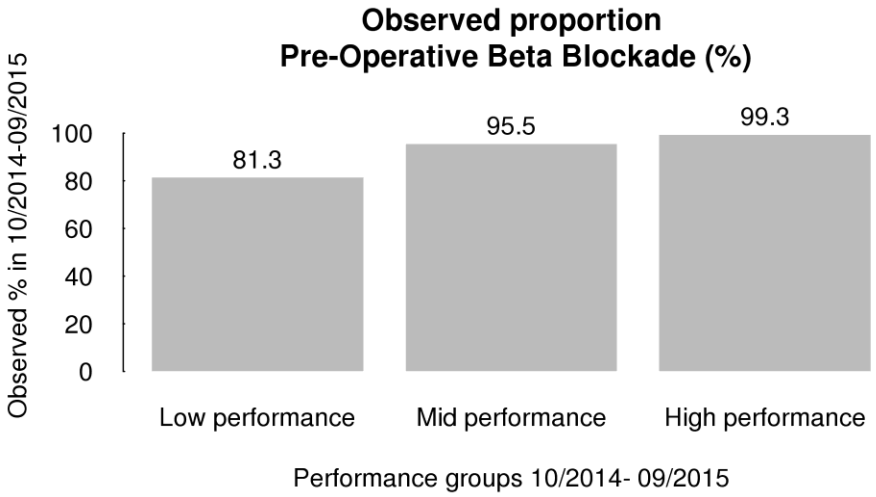
In addition, we tested the predictive validity of the measure. Predictive validity means that the results of this measure are predictive of future performance. We assessed the extent to which performance on this STS measure remains stable over time. In other words, does the measure at one point in time accurately predict performance at some later time?

The tests on validity used the concept of performance outliers to be more formally introduced in 2b5: Participants were labeled as "low performance" if the 95% exact binomial confidence interval of its event rate lies entirely below the population average (in other words, the upper bound of the 95% CI < population average). Participants were labeled as "high performance" if the 95% confidence interval lies entirely above 1. The remaining participants were labeled mid performance.

For each of the performance groups from the earlier period, we calculated the group specific measure proportions in the later period.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

STS participants deemed high performers by this measure have (on average) high rates of preoperative beta blockade. Thus, differences in performance were clinically meaningful as well as statistically significant. This is illustrated in the figure below using data from October 2014 to September 2015. Compared to participants who were deemed as having lower than average performance, those with better-than-average performance had higher rate of preoperative beta blockade (99.9% vs. 91.1%).

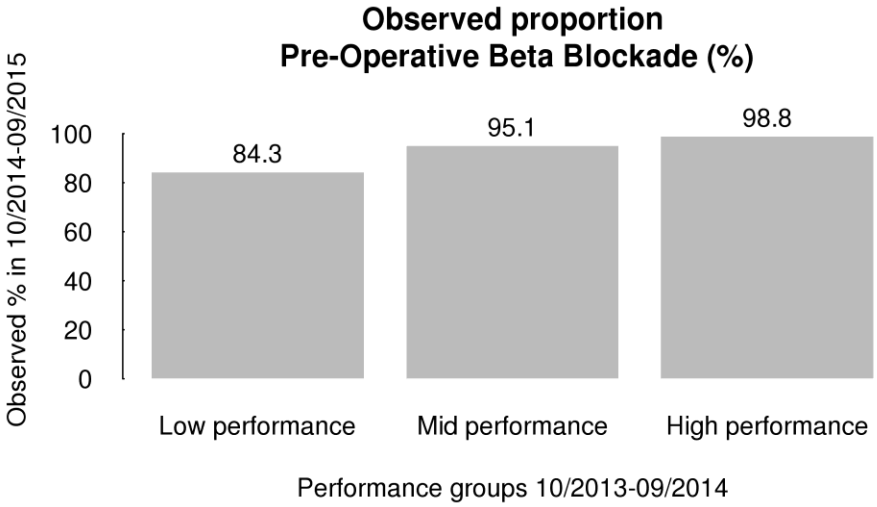


The predicted validity analysis was restricted to a sample of 1015 STS participants with patients receiving the measure in both time periods: October 2013 – September 2014 and October 2014 - September 2015. Among participants who were high performance centers in October 2013-September 2014, 77.0% of them were also high performers for October 2014 - September 2015. For comparison, only 12.0% of participants who were mid performers in October 2013-September 2014 became high performers in October 2014 - September 2015. Thus, participants who performed better than average in October 2013-September 2014 were over 6 times more likely to be identified as better performers in the next year. Similarly, participants who were low performance entities in the early year were more likely to remain low performers in the later year. 12 participants jumped from low to high performing status (or vice versa) between the two adjacent 12-month periods. Thus, a consumer may reasonably expect that a high or low performer will likely be the same or became average in the near future, and a mid-performer is likely to remain average.

Change in performance categories between two time periods

		10/2014 – 09/ 2015		
		Low performance	Mid performance	High performance
10/2013 - 09/2014	Low performance	130	56	9
	Mid performance	55	399	62
	High performance	3	67	234

For each of the performance groups in the earlier period, we also calculated its aggregated proportion of patients receiving the measure in the later period. The aggregated proportions in the later periods were 98.8%, 95.1%, and 84.3% for the high, mid and low performance groups from the earlier period.



2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show that the measure reflects the proportion of patients who were received preoperative beta blockade as designed, and that the past measure can be used to predict future performance. Together with face value, they support the validity of the measure.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

We excluded from the analysis cases for which preoperative beta blockade was contraindicated or if operative status was emergent or emergent salvage. We believe these are clinically appropriate exclusions and are necessary to make the measure a consistent performance measure for the comparison across participants. The exclusions are precisely defined and specified.

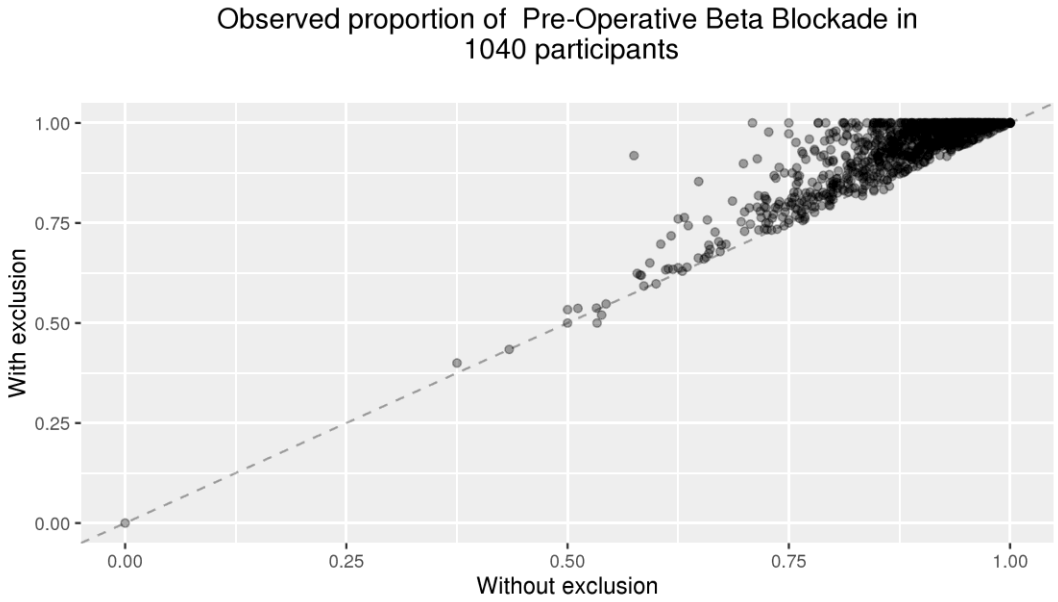
To show the impact of these exclusions, and how the measure would be distributed without them, we calculated and compared the distributions of the measure with and without the current exclusion criteria.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Distribution of participant-specific observed proportion of patients receiving the measure in October 2014 - September 2015 with and without the exclusion

Distribution	10/2014 – 09/2015 Observed proportion with exclusion	10/2014 – 09/2015 Observed proportion without exclusion
# Participant	1041	1040
# Operations	134689	147699
Mean	0.93	0.88
STD	0.093	0.093
IQR	0.082	0.099
0%	0.00	0.00
10%	0.82	0.76
20%	0.89	0.82
30%	0.93	0.86
40%	0.95	0.89
50%	0.97	0.91
60%	0.98	0.93
70%	0.99	0.94
80%	1.00	0.95
90%	1.00	0.97
100%	1.00	1.00
Midwest	297	297
Northeast	136	136
South	392	391
West	216	216
Low performance	197, 18.9%	204, 19.6%
Mid performance	538, 51.7%	599, 57.6%
High performance	306, 29.4%	237, 22.8%

Comparison of measure scores with and without the exclusion



The Spearman rank correlation of the measures with and without the exclusion is 0.74. The Pearson correlation is 0.89.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

For the measure to consistently quantify the quality per its definition, it is necessary to exclude cases if preoperative beta blockade was contraindicated or operative status was emergent or emergent salvage. It has an impact on the results for many participants, and the results would be distorted without these appropriate exclusions.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors *(e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)*

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The summary statistic provided is the participant's observed proportion of eligible patients who receive preoperative beta blockade.

The degree of uncertainty surrounding an STS participant's preoperative beta blockade measure estimate is indicated by the 95% exact binomial confidence interval (CI) of its observed proportion. Point estimates and CI's of the observed proportion for an individual STS participant are reported along with a comparison to the STS average proportion of the study time period. A performance category interpretation is also given to STS participants. **Since higher value indicates better performance**, an STS participant is designated as having higher/lower than average performance for the measure if the 95% CI lies entirely **above/below** the STS average. The remaining participants are labeled as not distinguishable from the STS average performance. For the simplicity of this report, we call the three groups 'high performance', 'low performance' and 'mid performance', respectively.

The method is equivalent to performing an exact binomial test with the null hypothesis that the participant has the same proportion of patients receiving the measure as the population average. Those with a test p-value smaller than 0.05 are the low and high performance groups.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

As shown in the table below, the proportion of STS participants performing better and worse than STS average has remained similar over the last two 12-month periods. On average, more than 50% of the participants have performance indistinguishable from the STS average, and the remaining participants have performed differently.

Distribution	10/2013 – 09/2014 Observed Proportion	10/2014 – 09/2015 Observed Proportion
# Participant	1057	1041
# Operations	134818	134689
Low performance	209, 19.8%	197, 18.9%
Mid performance	543, 51.4%	538, 51.7%
High performance	305, 28.9%	306, 29.4%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The statistical test and the construction of confidence interval are widely used and accepted. The participants identified as having performed differently from the average likely have true performance characteristics that are different. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the amount of outliers the measure detects.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS 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.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Due to great data quality, the source fields required by preoperative beta blockade had only 0.1% missing in the latest measure time window. We calculated the overall rate of missing as well as missing rates across all participants. In the implementation, missing data are imputed to "no". In addition, participants with greater than 5% missing data are excluded from the calculation of the measure.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

Overall 0.1% of data were missing. 99% of participants had missing rate of 4% or lower. Seven out of 1048 participants were not included because of having missing rates higher than 5%.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

The rates of missing data in the STS Adult Cardiac Surgery Database were very low and are getting lower. We therefore concluded that systematic missing data did not lead to bias in our measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. To what extent are the specified data elements available electronically in defined fields? *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Data Collection:
There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**
NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting STS Public Reporting Online: This measure is one of eleven component measures of the STS CABG Composite Score. Approximately 49.8% of STS Adult Cardiac Surgery Database participants are voluntarily enrolled in the STS public reporting program. STS Public Reporting Online: http://www.sts.org/quality-research-patient-safety/sts-public-reporting-online and Consumer Reports Health: www.ConsumerReports.org/hospitalratings Payment Program This is PQRS measure #44. The STS National Database was once again designated a Qualified Clinical Data Registry (QCDR) for PQRS reporting in 2016. STS reports this measure to CMS on behalf of all consenting surgeons. http://www.sts.org/quality-research-patient-safety/quality/physician-quality-reporting-system Quality Improvement with Benchmarking (external benchmarking to multiple organizations) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database Quality Improvement (Internal to the specific organization) The STS Adult Cardiac Surgery Database has more than 1,100 participants and nearly 6 million procedures http://www.sts.org/national-database/database-managers/adult-cardiac-surgery-database

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Please see table above.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

N/A

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Please see sections 1b.2 and 1b.4

In the table below we provide the overall trend over time of the measure performance at the patient level. The aggregate proportion of eligible patients receiving preoperative beta blockade was computed for each time period. Although measure performance was above 93% during the entire period, we can see a moderate increase on the proportion of patients receiving preoperative beta blockade over the last 4 years.

	10/2011 - 09/2012	10/2012 - 09/2013	10/2013 – 09/2014	10/2014-09/2015
All	93.25%	94.46%	94.84%	94.86%

Geographic area and number and percentage of accountable entities and patients included
Number of participants and operations by geographic regions, during the two last consecutive time periods, October 2013-September 2014 and October 2014-September 2015.

	10/2013 – 09/2014				10/2014 – 09/2015			
	Midwest	Northeast	South	West	Midwest	Northeast	South	West
# Participant	304	133	399	221	297	136	392	216
% Participant	28.8%	12.6%	37.7%	20.9%	28.5%	13.1%	37.7%	20.7%
# Operation	33307	21178	59253	21080	33129	21323	58907	21330
% Operation	24.6%	15.9%	43.9%	15.6%	24.5%	16.0%	43.7%	15.8%

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2015, 10% of participants, i.e., 107 facilities, were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0116 : Anti-Platelet Medication at Discharge
0117 : Beta Blockade at Discharge
0118 : Anti-Lipid Treatment Discharge

0119 : Risk-Adjusted Operative Mortality for CABG
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
0134 : Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** 0127_Preoperative_Beta_Blockade_Appendix_-_S.9-_1b.2-_1b.4-_Guidelines.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
• David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
• Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
• Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery

- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2004

Ad.3 **Month and Year of most recent revision:** 06, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0351
De.2. **Measure Title:** Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)
Co.1.1. **Measure Steward:** Agency for Healthcare Research and Quality
De.3. **Brief Description of Measure:** In-hospital deaths per 1,000 surgical discharges, among patients ages 18 through 89 years or obstetric patients, with serious treatable complications (shock/cardiac arrest, sepsis, pneumonia, deep vein thrombosis/ pulmonary embolism or gastrointestinal hemorrhage/acute ulcer). Includes metrics for the number of discharges for each type of complication. Excludes cases transferred to an acute care facility. A risk-adjusted rate is available. The risk-adjusted rate of PSI 04 relies on stratum-specific risk models. The stratum-specific models are combined to calculate an overall risk-adjusted rate.
1b.1. **Developer Rationale:** This indicator targets patients who are admitted for surgery who die following the development of a serious but treatable complication of care. Examples of such complications include: 1) shock or cardiac arrest, 2) sepsis, 3) pneumonia, 4) deep vein thrombosis or pulmonary embolism, and 5) gastrointestinal hemorrhage or acute ulcer. This indicator is fundamentally different than other PSIs, as it reflects the effectiveness of the hospital in rescuing a patient from complications versus preventing the underlying complications.

S.4. **Numerator Statement:** Number of deaths (DISP=20) among cases meeting the inclusion and exclusion rules for the denominator.
S.7. **Denominator Statement:** Surgical discharges, for patients ages 18 through 89 years or MDC 14 (pregnancy, childbirth, and puerperium), with all of the following:
• any-listed ICD-9-CM or ICD-10-PCS procedure codes for an operating room procedure; and
• the principal procedure occurring within 2 days of admission or an admission type of elective (ATYPE=3); and
• meet the inclusion and exclusion criteria for STRATUM_SHOCK (shock or cardiac arrest), STRATUM_SEPSIS (sepsis), STRATUM_PNEUMONIA (pneumonia), STRATUM_DVT (deep vein thrombosis or pulmonary embolism), or STRATUM_GI_HEM (gastrointestinal hemorrhage or acute ulcer)
STRATUM_SHOCK (shock or cardiac arrest)
• any secondary ICD-9-CM or ICD-10-CM diagnosis codes or any-listed ICD-9-CM or ICD-10-PCS procedure codes for shock or cardiac arrest
STRATUM_SEPSIS (sepsis)
• any secondary ICD-9-CM or ICD-10-CM diagnosis codes for sepsis.
STRATUM_PNEUMONIA (pneumonia)
• any secondary ICD-9-CM or ICD-10-CM diagnosis codes for pneumonia or pneumonitis.
STRATUM_DVT (deep vein thrombosis or pulmonary embolism)
• any secondary ICD-9-CM or ICD-10-CM diagnosis codes for deep vein thrombosis or pulmonary embolism.
STRATUM_GI_HEM (gastrointestinal hemorrhage or acute ulcer)
• any secondary ICD-9-CM or ICD-10-CM diagnosis codes for gastrointestinal hemorrhage or acute ulcer. Surgical discharges are defined by specific MS-DRG codes and ICD-9-CM/ICD-10-PCS codes indicating "major operating room procedures."
S.10. **Denominator Exclusions:** Exclude cases:
• transferred to an acute care facility (DISP = 2)
• with missing discharge disposition (DISP=missing), gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing)

De.1. **Measure Type:** Outcome
S.23. **Data Source:** Administrative claims
S.26. **Level of Analysis:** Facility

Original Endorsement Date: May 15, 2008 **Most Recent Endorsement Date:** Jan 31, 2012

IF this measure is included in a composite, NQF Composite#/title:
IF this measure is paired/grouped, NQF#/title:
De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? Not applicable

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This maintenance measure, last reviewed in 2011, measures in-hospital deaths per 1,000 surgical discharges, among patients ages 18 through 89 years or obstetric patients, with serious treatable complications (shock/cardiac arrest, sepsis, pneumonia, deep vein thrombosis/ pulmonary embolism or gastrointestinal hemorrhage/acute ulcer).
- The developer provided rationale to support the relationship between swift diagnosis and intervention by experienced, skilled nurses and a reduced probability of death from the specific complications stating that better quality nurses and higher nurse staffing levels have been shown to improve PSI 04 rates and that these rates are strongly associated with risk-adjusted mortality rates but not with complication rates.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**
☒ **The developer provided updated evidence for this measure:**

Updates:

- Although not required, the developer provided clinical practice guidelines that focus on early recognition and aggressive treatment of sepsis, antithrombotic therapy for venous thromboembolism, rapid diagnosis and stabilization of patients with gastrointestinal hemorrhage, and perioperative cardiovascular risk assessment and management.
- The developer also conducted an environmental scan of the literature on failure to rescue (FTR) in February 2016, and provided an updated summary of the body of evidence found through this recent search.

Exception to evidence: N/A

Questions for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat discussion and vote on Evidence?*
- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee believe there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm Health outcome (Box 1) →relationship between outcome and at least one healthcare action identified/supported by stated rationale (Box 2) → Pass

Preliminary rating for evidence: ☒ Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- In this submission, the developer notes a study that shows that due to improvements in care, between 1998 and 2007, FTR decreased by 6.05% per year (p<0.0001). However, the developer also notes that PSI 04 still captures approximately 43,000 deaths each year in the 34 states in the all-payer reference population. (Downey JR, Hernandez-Boussard T, Banka G, Morton JM. Is patient safety improving? National trends in patient safety indicators: 1998-2007. Health Serv Res. 2012;47(1 Pt 2):414-430.)
- The developer conducted all analyses using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID) from calendar years 2011-2013.
- The observed rate for death rate has declined from 118.0419 in 2011 to 116.3869 in 2013. Distribution of hospital performance for death rate in 2-year pooled data showed a median of 102.61 in 2011-2012 and 99.92 in 2012-2013

Table 1. Risk adjusted death rate per 1,000 surgical discharge, 2013

Patient/hospital characteristic	Risk-adjusted estimate
Total U.S.	**116.387**
Age 18-39	74.420
Age 40-64	100.445
Age 65 and over	133.431
Male	120.100
Female	111.841
First quartile (lowest income)	121.085
Fourth quartile (highest income)	112.644
Medicare	116.361
Medicaid	120.407
Uninsured/self-pay/no-charge	136.964
Northeast	122.821
Midwest	110.965
South	118.210
West	115.587

Disparities

- The developer provided data (see Table 1 above) among 182,512 surgical patients with serious treatable conditions, older patients, men, those from lower income communities, those with Medicare, Medicaid or uninsured, and those treated in the Northeast were at greater risk of death (after controlling for a variety of clinical risk factors) than younger patients, women, those from higher income communities, those with private insurance and those treated in the Midwest or West.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *What does the data tell us about disparities for this condition?*

Preliminary rating for opportunity for improvement: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- Why exclude patients transferred to acute care? (can't track them in HCUP)
- Only reliable in larger hospitals (shrinkage adjustments used)
- consistent correlations with teaching status, advanced nursing staff, nursing ratio, high-tech (? what this means)
- Complex risk-adjustment per complication.
- this is an outcome measure of the number of deaths in surgical patients who develop serious treatable conditions postoperatively (shock, pneumonia, DVT, hemorrhage). It is linked to the quality of nursing care in evidence provided by the developer.
- The measure meets this criterion. In addition to the original supporting evidence, the developer provided clinical practice guidelines and an environmental scan of literature supporting early recognition and treatment and failure to rescue, both related to the outcome tracked by this measure.

1b.

- there still is a significant gap- 43,000 deaths per year in 34 measured states in an all payor dataset.
- older men and those with less well-paying or no insurance have higher mortality than the median rate for this measure.
- Although performance appears to be improving, the remaining gap is large enough to warrant a national performance measure. The results to appear to vary by some of the population subgroups provided, although I found myself desiring more context to understand this amount of variation vs. performance variation between/among facilities.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): Administrative claims

- All analyses were completed using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), 2011-2013.

Specifications:

- The measure is specified as a facility-level measure for the hospital/acute care setting.
- The numerator includes the number of deaths (DISP=20) among cases meeting the inclusion and exclusion rules for the denominator.
- The denominator includes the total number of surgical discharges, defined by specific MS-DRG codes and ICD-9-CM/ICD-10-PCS codes indicating "major operating room procedures," for patients ages 18 through 89 years or MDC 14 (pregnancy, childbirth, and puerperium), with all of the following:
  - any-listed ICD-9-CM or ICD-10-PCS procedure codes for an operating room procedure; and
  - principal procedure occurring within 2 days of admission or an admission type of elective (ATYPE=3); and
  - meet the inclusion and exclusion criteria for STRATUM_SHOCK (shock or cardiac arrest), STRATUM_SEPSIS (sepsis), STRATUM_PNEUMONIA (pneumonia), STRATUM_DVT (deep vein thrombosis or pulmonary embolism), or STRATUM_GI_HEM (gastrointestinal hemorrhage or acute ulcer)
- This outcome measure is risk adjusted, using a statistical risk model for each stratum.
- ICD-9 and ICD-10 codes, and a conversion table are provided.
- Denominator exclusions are cases that were transferred to an acute care facility (DISP = 2) and cases with missing discharge disposition (DISP=missing), gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing)
- The developer notes that annual updates are completed for all measures according to standard protocol. In addition, approximately every two years, AHRQ updates the risk-adjustment parameter estimates based on the

most recent year of data (i.e., the most current reference population possible). Therefore, since the last update, there have been some changes to the specifications in forthcoming 2016 version 6.0 specifications since the previously endorsed version (v4.4).

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Are 2016 specification changes reasonable for this measure?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- In the previous review, the developer conducted empirical analysis using AHRQ 2007 State Inpatient Databases (SID) with 4,000 hospitals and 30 million discharges, which the Committee found to be acceptable.

Describe any updates to testing

- Updates have been made to the reliability testing since the last review. In this submission, additional performance measure score reliability testing has been conducted using 2-year pooled data (2012-2013) from the HCUP State Inpatient Databases (SID). These data were obtained from 3,199 hospitals.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**
Reliability testing performed with the data source and level of analysis indicated ☒ **Yes** ☐ **No**

Method(s) of reliability testing:

- The developer completed *performance measure score* testing using a signal-to-noise analysis, appropriate for this measure, which assesses differences in performance between hospitals ("the signal") to stability within hospitals (random measurement error or "the noise").
- Hospital size is used as a weight when calculating an overall reliability estimate. Weighting reduces the influence of hospitals that have less reliable rates due to a smaller number of patients at risk (small denominators).

Results of reliability testing

- Performance score testing results are provided by deciles, based on hospital size, for 319 and 320 hospitals from 2012-2013 HCUP SID data.
  - Overall signal-to-noise ratio of 0.60 for 3,199 hospitals with an overall average of 115 discharges per year,
    - a range from 0.4738 to 0.7765 for 1,280 hospitals with > 82 average discharges per year, and
    - a range from 0.0579 to 0.3954 for 1,919 hospitals with < 82 average discharges per year.
- The signal-to-noise method results in a reliability statistic that ranges from 0 to 1 for each facility. A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences between hospital performance. A value of 0.7 is often regarded as a minimum acceptable reliability value. From the data presented only hospitals in the decile representing the largest hospitals (437.5 average discharges) have reliability above 0.7.
- The developer notes that " the AHRQ QI program generally considers ratios between 0.4 – 0.8 as acceptable. It is rare to achieve reliability above 0.8."

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) → empiric reliability testing (Box 2) →Testing of the measure score (Box 4) → appropriate method (Box 5) → testing results (Box 6) →6b moderate certainty

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

- As a part of the original endorsement process in 2007, the developer conducted empirical validity testing based on Medicare inpatient fee-for-service claims for general surgical admissions from July 1,1999 through June 30, 2000 using data including 1,467 hospitals and 403,679 Medicare beneficiaries between 65 and 90 years of age. The developer assessed construct validity by estimating logic models using detailed patient characteristics and 5 hospital characteristics shown to be associated with better quality of care in previous studies.
- Omega statistics were also evaluated in these analyses. The omega statistic represents the ratio of the squared sum of the log odds for patient characteristics at the discharge-level variables divided by the corresponding quantity for hospital-level variables. Lower omega ratios may be indicative of a more desirable quality indicator. The omega ratio summarizing the contribution of patient characteristics at the discharge-level versus hospital-level variables was 57, compared to a ratio of 189 for the overall risk adjusted surgical mortality rate and 128 for NQF #0352.
- The developer also evaluated face validity using a structured panel review process that was based on the RAND appropriateness method, a modified Delphi process known as nominal group technique; the panel was convened in 2002 and was comprised of 7 multispecialty members.

Describe any updates to validity testing

- Empirical validity testing has been updated since the last review, using HCUP SID reference data from 2012-2013. The analyses use a broader definition of teaching status. The developer states that teaching hospitals had higher unadjusted PSI 04 rates, but lower adjusted PSI 04 rates relative to nonteaching hospitals but that this effect was less pronounced with the more inclusive definition of teaching hospitals (and all-payer data instead of Medicare data).

SUMMARY OF TESTING

Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☐ **Face validity only**
☒ **Empirical validity testing of the measure score**

Validity testing method:

- The additional empirical validity testing was conducted by correlating the measure score to hospital characteristics, including hospital teaching status. The developer used all-payer data from 34 states in 2012-2013 to confirm the association between hospital teaching status and PSI 04.

Validity testing results:

- Risk adjusted death rates are reported as overall reference population rates and as a distribution of performance for the measure.
- Empirical validity testing generally demonstrated favorable correlations between the measure score and hospital teaching status. As hypothesized, teaching hospitals and high volume PCI hospitals have lower PSI 04 rates. The following table shows odds ratio and p value for structural measures of quality when all patient characteristics are included in the model with only one hospital characteristic at a time:

Table 2.

Hospital characteristic	Marginal analysis Odds ratio (p-value)	Partial analysis Odds ratio (p-value)	Interpretation
Teaching hospital	0.807 (p<0.0001)	0.852 (p<0.0001)	Protective
High tech hospital	0.924 (p<0.005)	1.049 (NS)	Protective
Large hospital	0.872 (p<0.0001)	0.917 (p<0.01)	Protective
Less well-staffed hospital	1.108 (p<0.005)	1.044 (NS)	Not protective
Better nursing skill mix	0.832 (p<0.0001)	0.870 (p<0.0001)	Protective

- The developer notes that as summarized in the Evidence Form, numerous studies have linked failure to rescue measures, including PSI 04, to structure and process measures. Multiple studies have found lower FTR rates in hospitals with higher nurse-to-bed ratios, better nurse skill mix ratios, and better US-trained nurse ratios. Higher hospital volume was associated with lower FTR rates in at least 6 studies. In addition, studies have found that hospitals with the highest patient satisfaction scores and hospitals with better compliance with NQF Safe Practices had lower risk adjusted odds of FTR.
- The developer notes that although they did not conduct new face validity, they anticipate the results of face validity to be similar if a new panel was convened because the characteristics of these events, treatment and prevention of approaches have not changed substantially since 2002.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- The developer conducted an empirical evaluation of exclusions using the 2013 data from 34 states to examine the percent of potential denominator cases excluded.
- The stratum-specific exclusions are meant to exclude cases for which the "complication" was actually the principal reason for admission or the primary indication for surgery. Some exclusions (e.g. immunocompromised state) are intended to exclude patients for whom death may be the expected outcome (i.e., less preventable). The table below summarizes the highly frequent exclusions.

	Denominator		
	Exclusion Count	After Exclusions	% Change
No Exclusion		329,827	
Transfers to acute care facility	9,889	319,827	3.0%
Pneumonia: Dx of immunocompromised	8,696	321,020	2.6%
Pneumonia: MDC 4	9,006	320,710	2.7%

Sepsis: Principal dx of Septicemia	45,202	284,514	13.7%
Sepsis: Principal dx of Infection	14,982	314,734	4.5%
Shock/Cardiac Arrest: MDC 4 or 5	27,819	301,897	8.4%
GI hemorrhage/Acute ulcer: MDC 6 0r 7	14,134	315,582	4.3%

- All patients transferred to other hospitals must be excluded from the analysis because the relevant outcome of these patients (i.e., dead or alive at the time of discharge from the acute inpatient setting) cannot be ascertained without social security numbers or other data elements to support linkage.
- Based on Needleman et al.'s analysis, AHRQ does NOT exclude patients with complications that were present on admission (i.e., upon transfer from another hospital, emergency department, or ambulatory surgery center).

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Risk adjustment summary

- The developer used a statistical risk model with 362 risk factors.
- Factors in the risk adjustment model were considered as a standard set of covariates grouped into four categories: demographics, severity of illness (Major Diagnostic Categories [MDCs]), Modified Diagnostic Related Groups [MDRGs]), comorbidities (AHRQ Comorbidity Software categories), and transfer-in status.
- The risk-adjusted rate for the overall PSI 04 is calculated as the observed to expected ratio multiplied by the reference population rate (the observed and expected values are summed across five categories of PSI 04 risk).
- Only those covariates present in at least 30 records for that PSI 04 strata are retained. A parsimonious model was identified using backward stepwise selection with bootstrapping.
- The analysis evaluates performance of the risk adjustment model(s) with respect to in-hospital death. The developer used c-statistic to measure how well the risk adjustment model distinguishes events from non-events.
- There are five distinct risk models (one for each type of complication or stratum).These five models currently have c-statistics ranging from 0.726 to 0.860. The c-statistic for the overall PSI 04 model is 0.829 in the 2013 HCUP data and it represents the overall performance of all five models.
- The developer notes that SDS variables were not included in the risk adjustment because there was no evidence or causal model to suggest that socioeconomic factors are associated with in-hospital death following serious surgical complications independent of quality of care, or are mediated by pre-hospital care (which may not fall within the proper realm of hospital accountability).

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the appropriate risk adjustment variables being used?*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in the risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- According to the developer, over all hospitals, using smoothed rates, this indicator has limited discrimination for identifying low or high performing hospitals; 16% of hospitals can be classified as better or worse than the threshold and 27% as better or worse than the benchmark, based on conventional statistical criteria. As hospital size increases, the discrimination also increases.
- For this analysis, the developer notes that "benchmark" refers to the smoothed indicator rate based on the 20th percentile of the reference population (i.e., 20% of hospitals have a lower mortality rate or better performance). "Threshold" refers to the indicator rate based on the 80th percentile (i.e., 80% have lower mortality or better performance).
- Assuming an underlying Gamma distribution for the smoothed rates of the measure, the benchmark and

threshold values are identified using population reference rates and signal variances computed from the entire AHRQ QI POA Reference Population.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- N/A

2b7. Missing Data

- PSI 04 excludes cases with missing discharge disposition, age, sex, discharge quarter, discharge year, and principal diagnosis. These variables are required for indicator construction and are required of all hospital discharge records.
- For these variables, frequencies of missing data are typically less than 1% of the state database. The developer notes that it is unlikely that bias would occur from such a low frequency of missing data.

Guidance from the Validity Algorithm: Specifications consistent with the evidence (Box 1) → Potential threats to validity addressed (Box 2) →Empirical validity testing performed (Box 3) →testing of measure score (Box 6) →Method was appropriate (Box 7) → Box 8 (high certainty) – High
Note that the body of literature has validated this measure.
Preliminary rating for validity: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.

- very clear documentation of data elements, defined using administrative and billing codes (ICD-9 and 10).
- It appears to me that this measure is only intended to be used in the HCUP SID dataset, which is publically available. It can't be used to identify specific hospitals for quality improvement efforts, though, I believe.
- In our organization's experience with the AHRQ QI's, consistent implementation is greatly enhanced by the fact that AHRQ providers the code/software, limiting the number of judgement calls that need to be made.

2a2.

- the developer tested the 'signal to noise' ratio for the measure, and this was acceptable (>0.7) for only the largest hospitals (>436 discharges). The developer feels that 0.4-0.8 is acceptable, which would include all hospitals in the dataset.
- I agree with the algorithm results on the measure worksheet. The testing was performed on a large data set and the results demonstrate moderate reliability.

2b1.

- I think the specifications align with the evidence presented.

2b2.

- This has been done in several ways. In prior applications, the developers conducted face validity with a panel of experts using the Delphi method. They also looked at performance of the measure using Medicare data to look at 5 other elements of hospital structure associated with better outcomes. They also generated an 'omega statistic', comparing the ratio of discharge to hospital-level odds ratios, which was acceptable compared to other measures. They recently updated validity testing by examining how teaching hospitals fared with and without risk adjustment and found reasonable results.
- Testing was performed with a large and diverse data set.

2b3.

- exclusions are reasonable and the rate of exclusions is presented. risk adjustment is intensive; goodness of fit statistics are high for risk adjusted models testing for event vs non event.
- SES is specifically not adjusted for.
- the ability to discriminate (meaningful differences) is moderate. only a quarter of hospitals can be distinguished as above or below the benchmark. This is improved when looking at larger hospitals only."
- I have no concerns with the validity of this measure

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer noted the following:

- All data elements are defined fields in electronic claims. The data are coded by someone other than the individual obtaining the original information. (e.g., DRG, ICD-9 codes on claims)
- The indicator is based on readily available administrative billing and claims data and U.S. Census data, thus is very feasible.
- The AHRQ QI software has been publicly available at no cost since 2001.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

- The developer states that all data elements are generated during the claims generation process, making this a feasible measure to report. It appears that AHRQ software is required to classify the claims diagnoses into the data fields used to construct the measure. This is an extra step but is not terribly burdensome.
- We calculate this measure and know several other organizations that do as well and all agree it's very straightforward to use the AHRQ QI software to generate measure results.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

See section 4.1 for list of programs.

Improvement results

- See Table 1 in response to question 1b.2 (also included in supplemental materials)
- The developer noted that the observed that PSI 04 rates have been relatively stable from 2011-2013 in the AHRQ QI POA Reference Population data (116-118 deaths per 1,000 patients with perioperative or postoperative complications).
- An earlier study of administrative data showed a decrease by 6.05% per year from the years 1998 to 2007. ($p<0.0001$) (Downey et al., 2012).

Unexpected findings (positive or negative) during implementation

- No evidence has been identified suggesting unintended consequences or findings for this measure.

Potential harms:

- No evidence has been identified suggesting unintended consequences for this measure.

Feedback : N/A

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

- This measure has been used in dozens of programs listed by the developer. It was associated with a 6.5% annual rate improvement initially, though the rate has stabilized in recent years. I can think of no unintended consequences.
- The measure is widely used.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0352 : Failure to Rescue In-Hospital Mortality (risk adjusted)
- 0353 : Failure to Rescue 30-Day Mortality (risk adjusted)

Harmonization

- AHRQ response to harmonization in current submission is linked here.
- During the previous evaluation of this measure the Committee discussed that these measures, while conceptually similar, have different aims, i.e., capture of avoidable complications vs. failure to rescue (NQF #0352 and #0353). Overall the Committee agreed that these measures have different objectives and are complementary.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 0351

Measure Title: Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 5/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Health outcome: In-hospital mortality

☐ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

This indicator targets patients who are admitted for surgery (defined by specific MS-DRGs and an ICD-9-CM/ICD-10-PCS code for an operating room procedure, and a principal procedure within 2 days of admission OR admission type of elective) who die following the development of a serious, but treatable complication of care. Complications included in PSI 04 are as follows: 1) shock or cardiac arrest, 2) sepsis, 3) pneumonia or pneumonitis, 4) deep vein thrombosis or pulmonary embolism, and 5) gastrointestinal hemorrhage/acute ulcer. Evidence supports the concept that swift diagnosis and intervention by experienced, skilled nurses lowers the probability of death from these specific complications. This concept has been described as "failure to rescue (FTR)" to reflect the benefits of recognizing and responding skillfully to early signs of patient deterioration. Both better quality nurses and higher nurse staffing levels have been shown to improve PSI 04 rates. PSI 04 rates have been shown to be strongly associated with risk-adjusted mortality rates, as expected, but not with complication rates.[1]

Please note: Relevant text regarding evidence of the health outcome for previous submissions to NQF is included below in black.

Mortality is a frequent outcome among patients with serious treatable complications

Silber and colleagues have published a series of studies establishing the construct validity of failure to rescue rates through their associations with hospital characteristics and other measures of hospital performance. Among patients admitted for cholecystectomy and transurethral prostatectomy, failure to rescue was independent of severity of illness at admission, but was significantly associated with the presence of surgical housestaff and a lower percentage of board-certified anesthesiologists. The adverse occurrence rate was independent of this hospital characteristic. In a larger sample of 74,647 patients who underwent general surgical procedures in 1991-92, lower failure to rescue rates were found at hospitals with high ratios of registered nurses to beds. Failure rates were strongly associated with risk adjusted mortality rates, as expected, but not with complication rates. Finally, among 16,673 patients admitted for coronary artery bypass surgery, failure rates were lower (whereas complication rates were higher) at hospitals with magnetic resonance imaging facilities, bone marrow transplantation units, or approved residency training programs.

More recently, Needleman and Buerhaus confirmed that higher registered nurse

staffing (RN hours/adjusted patient day) and better nursing skill mix (RN hours/licensed nurse hours) were consistently associated with lower failure to rescue rates among major surgery patients from 799 hospitals in 11 states in 1997, even using administrative data to define complications. An increase from the 25th to the 75th percentile on these two measures of staffing was associated with 5.9% (95% CI, 1.5% to 10.2%) and 3.9% (95% CI, -1.1% to 8.8%) decreases, respectively, in the rate of failure-to-rescue among major surgery patients. These associations were inconsistent among medical patients, in that nursing skill mix was associated with the failure-to-rescue rate (rate ratio 0.81, 95% CI 0.66-1.00) but aggregate registered nurse staffing was not (rate ratio 1.00, 95% CI 0.99-1.01). An increase from the 25th to the 75th percentile on nursing skill mix was associated with a 2.5% (95% CI, 0.0% to 5.0%) decrease in the failure-to-rescue rate among medical patients.

Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. Med Care 1992;30(7):615-29.

Silber J, Rosenbaum P, Ross R. Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics? J Am Stat Assoc 1995;90:7-18.

Silber JH, Rosenbaum PR, Williams SV, Ross RN, Schwartz JS. The relationship between choice of outcome measure and hospital rank in general surgical procedures: Implications for quality assessment. Int J Qual Health Care 1997;9(3):193-200.

Needleman J, Buerhaus PI, Mattke S, Stewart M, Zelevinsky K. Nurse Staffing and Patient Outcomes in Hospitals. Boston MA: Health Resources and Services Administration; 2001 February 28. Report No.:230-99-0021.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

Please note that this is an outcome measure, so a systematic review of the body of evidence that supports the performance measure is not required. However, information is provided in 1a.4.1 and 1a.8 below, to provide additional context and support for the measure. Specifically, there are several high-quality systematic reviews and clinical practice guidelines that focus on early recognition and aggressive treatment of sepsis, antithrombotic therapy for venous thromboembolism, rapid diagnosis and stabilization of patients with gastrointestinal hemorrhage, and perioperative cardiovascular risk assessment and management. These practice guidelines reflect the process-outcome pathways that are facilitated by highly skilled, well-functioning multidisciplinary teams.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1. Dellinger RP, Levy MM, et al. Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Crit Care Med. 2013 Feb;41(2):580-637. doi: 10.1097/CCM.0b013e31827e83af.
2. Fleisher LA, Fleischmann KE, Auerbach AD, et al. 2014 ACC/AHA guideline on perioperative cardiovascular evaluation and management of patients undergoing noncardiac surgery: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014; 130:2215.
3. Kearon C, Akl EA, Ornelas J, et al Antithrombotic therapy for vte disease: chest guideline and expert panel report. *Chest* 2016;149(2):315-352. doi:10.1016/j.chest.2015.11.026.

4. Konstantinides SV, Torbicki A, et. al. Task Force for the Diagnosis and Management of Acute Pulmonary Embolism of the European Society of Cardiology (ESC).2014 ESC guidelines on the diagnosis and management of acute pulmonary embolism Eur Heart J. 2014 Nov 14;35(43):3033-69, 3069a-3069k. doi: 10.1093/eurheartj/ehu283. Epub 2014 Aug 29.
5. Laine L, Jensen DM. Management of patients with ulcer bleeding. Am J Gastroenterol. 2012 Mar;107(3):345-60. [133 references] http://www.guideline.gov/content.aspx?id=38023&search=gi+hemorrhage, accessed February 23, 2016.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Not applicable

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Not applicable

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Not applicable

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

Not applicable

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

Not applicable

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

Not applicable

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

Not applicable

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

Not applicable

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Not applicable

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Not applicable

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Not applicable

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

Not applicable

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Not applicable

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Not applicable

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Not applicable

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

Not applicable

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Not applicable

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not applicable

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Not applicable

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Not applicable

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

Not applicable

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

Formal environmental scans of the literature, including routine PubMed searches are performed to continually update evidence. The current evidence review results presented below constitute the most recent update, conducted in February 2016. Search terms included relevant MeSH terms (serious treatable complications, PSI 4 or PSI 04). The search was limited to English language publications. For completeness we also tested more inclusive search strings. Below we have provided a summary of the most up-to-date evidence.

1a.8.2. Provide the citation and summary for each piece of evidence.

Summary

The concept of failure to rescue (FTR), initially developed by Silber, has been widely studied.[2,3] Silber et al. published a series of studies linking risk adjusted mortality rates following complications to various physician factors, including residents and surgical housestaff, board-certified anesthesiologists, board certified surgeons, registered nurse staffing ratios.[2-4] Needleman and Buerhaus adapted Silber's concept of FTR for use with administrative data, targeting nurse sensitive measures by narrowing the number of complications included.[5,6] The Buerhaus measure was further modified and operationalized by AHRQ into the current PSI 04 specifications (first endorsed by NQF in 2008).[6] Ongoing research continues to confirm the association with nurse staffing, nurse skill mix, and other structural measures and processes of care (as outlined as follows) with PSI 04 and the similar FTR measure.[4,5,7-19] Johnston and colleagues recently published a systematic review of 42 articles to "identify the factors that contribute to high FTR rates and delayed escalation of care, and… summarize outcomes of interventions aimed at decreasing the rates of FTR and improving escalation of care."[3]

Structural measures: Nursing

FTR rates are associated with many hospital level measures of high-quality care. One of the most studied areas is the association of nursing characteristics and PSI 04 (including the variant measure for FTR).[4,5,8-19] As stated above, PSI 04 has special relevance to nursing, as nurses are on the front line of patient care, with key roles in prevention, surveillance, early detection, and intervention. Multiple studies have found lower FTR rates in hospitals with higher nurse-to-bed ratios, better nurse skill mix ratios, and higher US-trained nurse ratios.[5,9,11-17,19-24]

The most recent studies include those by Unruh et al, Park et al., and Blegen et al.[22,24,25] Using latent growth curve models, Unruh et al. found that increases in RN full time equivalents were associated with decreases in PSI 04 in 124 Florida hospitals over 9 years.[25] In addition to nurse staffing, Park et al. examined patient turnover rates in non-ICUs (56.1%) and ICUs (45.4%) in 42 teaching hospitals in 2005.[24] In general, they found that more RN hours per patient day were associated with lower rates of PSI 04, controlling for non-RN staffing and hospital characteristics. Patient turnover was not related to PSI 04 rates in either non-ICUs or ICUs; however, the association between RN staffing and PSI 04 in non-ICU settings differed significantly depending on the level of patient turnover. Specifically, when patient turnover rates increased from 48.6% (25th percentile) to 60.7% (75th percentile), the effect of RN staffing on PSI 04 was reduced by 11.5%.[24] Using 21 of the same teaching hospitals, which agreed to provide detailed data on nurse education over 4 quarters, Blegen et al. found that the proportion of baccalaureate-educated nurses was inversely associated with PSI 04 ($r = -0.399$; $p < 0.05$).[22]

Magnet designation by the American Nurses Credentialing Center is a related concept that captures nursing empowerment and excellence. Mills and Gillespie matched 80 Magnet hospitals with 80 non-Magnet hospitals in the 2001-2005 Nationwide Inpatient Samples on 12 hospital characteristics and reported no significant difference in PSI 04 rates between matched hospitals.[26] McHugh et al. focused on four states (CA, FL, NJ, PA) in 2006-2007, and linked state hospital discharge data with detailed data on nurse

characteristics and work environments, from surveys of over 100,000 registered nurses.[27] A composite measure of nursing, estimated as the likelihood of a hospital being Magnet credentialed as a function of nursing factors, was significantly associated with lower odds of failure-to-rescue (OR 0.48, 95% CI 0.37-0.63). The Practice Environment Scale of the Nursing Work Index was the most important component of that composite. When an indicator for Magnet status was added to this model, the composite nursing measure was still significantly associated with failure-to-rescue (OR 0.57, 95% CI 0.41-0.79) and the Magnet effect approached significance (OR 0.88, 95% CI 0.77-1.01), indicating that most but not all of the Magnet effect can be explained by other aspects of the nursing work environment. Kutney-Lee et al. extended these findings by showing that between 1999 and 2007, 11 newly Magnet recognized hospitals in Pennsylvania reduced their FTR rates by 6.1 per 1000 patients (P=0.02) compared with 125 non-Magnet comparison hospitals.[28] Finally, Friese et al. used Medicare data on patients hospitalized for coronary artery bypass graft surgery, colectomy, or lower extremity bypass in 1998–2010 to show that patients treated in Magnet hospitals were 8.6% (95% CI: 0.88, 0.95) less likely to die from FTR than patients treated in non-Magnet hospitals.[29]

One recent study found mixed effects with respect to the impact of regulations designed to improve nurse staffing on PSI 04 rates in California. In this study, Mark et al. divided CA hospitals into quartiles based on their pre-regulation nurse staffing levels for medical-surgical and pediatric services (Quartile 1 = lower RN staffing).[30] Using difference-in-differences Poisson fixed effects models to compare California with 12 other states, they found that PSI 04 decreased significantly more in California Quartile 1 hospitals than in comparison state hospitals in both the immediate pre-regulation and post-regulation periods (by 37.1% and 30.7%, respectively, $p < .05$), while RN staffing improved by 27.3% and 35.0%, respectively. However, PSI 04 also decreased significantly more in California Quartile 4 hospitals than in comparison state hospitals in the post-regulation period (by 32.9%, $p < .05$), even though these hospitals only improved their RN staffing by 15.1%.

Structural Measures: Physician staffing

As with nursing, higher resident-to-bed ratios and higher case volume are associated with lower FTR rates. Confirmatory findings have been reported in general surgery, orthopedics, vascular surgery, and cardiovascular surgery.[11,20,31-33] The use of resident housestaff is hypothesized to increase ability to identify and act upon complications early.

Most recently, a 2013 retrospective cohort analysis of the National Surgical Quality Improvement Program (NSQIP) data (2005-2009) found that resident trainee participation in complex, oncologic surgery was associated with significantly higher rates of 30-day postoperative complications in NSQIP-participating hospitals, but lower 30-day mortality and lower FTR among patients suffering complications (5.9 vs. 7.6%, AOR 0.79, 95% CI 0.68-0.90).[34] Ferraris et al (2014) analyzed NSQIP data from 200 hospitals and found that cases with surgical resident involvement had increased operative morbidity (11.4% vs 7.8% with attending only; $P < .001$) and prolonged operative time (127 minutes vs 93 minutes for attending only; $P < .001$), but more favorable FTR (9.4% vs 12.4% for attending alone; $P < .001$).[35] The most serious complications occurred 5 to 10 days before death, suggesting that there is a window for intervention to rescue patients with early aggressive treatments for the sentinel complication. Among 1,056,865 CABG patients in the 1998-2007 Nationwide Inpatient Sample, teaching hospitals had 14% lower FTR rates (OR 0.86, 95% CI 0.79-0.93) than non-teaching hospitals, despite higher complication rates (OR 1.023, 95% CI 1.00-1.05), in the first quarter of the academic cycle.[36] Finally, Navathe et al. (2013) found no differences in FTR in teaching hospitals when comparing rates prior to and after ACGME duty hour reform.[37]

Patient clinical and sociodemographic factors

Risk for FTR varies by procedure and complication type.[35,38] In the NSQIP database, timing of complication also impacted mortality rate, with early cardiac arrest and unplanned intubation associated with lower risk-adjusted mortality (HR 0.59, CI 0.51-0.68; HR 0.38, CI 0.33-0.43, respectively).[39] Late pneumonia was associated with higher observed mortality but not adjusted mortality. Ferris et al. reported that more than two-thirds of patients with failure to rescue have multiple complications.[35] Silber et al. found that blacks had higher failure-to-rescue rates (6.1% vs. 5.1%, P<0.001) than whites, matched on age, sex year, state and procedure in Medicare data. However, when preoperative medical risk factors were added to this matching algorithm, there was no significant racial difference in FTR. [40] Similarly, Black patients had higher odds of observed FTR than white patients (1.14, 95% CI 1.07-1.23) in four states, although this risk was not significant when adjusted by SES, and the white patients had higher risk than black patients when adding other patient characteristics to the model.[41] The study identified a modest but significant interaction between race and nurse staffing for FTR.[40] Each additional patient in a nurse's workload raised the odds of FTR for black patients by a factor of 1.10 (95% CI 1.02-1.18) and 1.04 for white patients (95% CI 1.01-1.06).[41] Among Medicare patients undergoing cancer surgery, patients in the lowest SES quintile had the highest FTR rates (26.7% vs 23.2%, P = .007), and hospitals treating larger portions of SES also had higher rates. However, these disparities did not remain after adjustment for hospital effects.[42]

Other hospital level factors

Higher hospital volume was associated with lower FTR rates in at least 6 studies.[5,31,33,43-46] Across three high complexity

cancer procedures, volume was more strongly associated with FTR rates than complication rates (lowest volume quintile odds ratio 1.17, 95% CI 1.02-1.33 for complications vs. 2.89, 95% CI 2.40-3.48 for FTR, compared to the highest volume quintile).[31] Among 119,434 Medicare cardiovascular surgery patients, FTR was consistently lower among hospitals with higher procedural volume across procedure types.[33] Similar results have been found for hepatic and hepato-pancreatico-biliary procedures, ovarian cancer resection, and surgical oncology patients.[43-46] In one study, NCI cancer centers had lower FTR rates than other hospitals among surgical oncology patients (OR 0.68, 95% CI 047-0.97).[47] Among cardiac surgery patients at 17 hospitals, hospitals with lower FTR rates had longer postoperative and intensive care unit stays after the index operation (2 to 3 days; p<0.001).[48] Wakeam et al. (2014) found that higher burden, as classified by the Safety Net hospital categories, was associated with higher odds of adjusted FTR for High Burden Hospitals (OR, 1.35; 95% CI, 1.19-1.53; P < .001) and Moderate Burden Hospitals (OR, 1.15; 95% CI, 1.05-1.27; P = .005) compared with Low Burden Hospitals.[49]

Hospital processes of care and outcomes measures

FTR has been studied in relationship to other quality measures. Sacks et al (2015) linked Medicare claims with NSQIP and Hospital Compare data on reported patient satisfaction. Hospitals in the highest quartile of Hospital Consumer Assessment of Healthcare Providers and Systems scores had significantly lower risk-adjusted odds of FTR (OR = 0.82, 95% CI, 0.70-0.96).[50] Supporting the preventability of FTR events, Brooke, et al. (2012) found that hospitals that complied fully with the 27 National Quality Forum (NQF) safe practices had an increased likelihood of diagnosing a complication after any of six high-risk operations (odds ratio [OR], 1.13; 95% confidence interval [CI], 1.03-1.25), but had a decreased likelihood of failure to rescue (OR, 0.82; 95% CI, 0.71-0.96), and a decreased odds of mortality (OR, 0.80; 95% CI, 0.71-0.91).[51]

Several studies have examined interventions to reduce FTR. In a single site study, an intervention including crew resource management and checklist implementation was associated with a reduction in FTR from 25% to 12% (P=0.03).[52] Two studies demonstrate lower FTR rates with more available technology.[4,53] In surgical oncology, NCI cancer center designation has also been found to be associated with reduced FTR rates following cancer surgery [47,54] Sheetz and colleagues evaluated whether increased hospital care intensity (HCI) is associated with improved outcomes following seven major cardiovascular, orthopedic, or general surgical operations in the Medicare population. High-HCI hospitals had greater rates of major complications than low-HCI centers (risk ratio, 1.04; 95% CI, 1.03-1.05) and there was a decrease in failure to rescue at high compared with low-HCI hospitals (risk ratio, 0.95; 95% CI, 0.94-0.97). Using multilevel-models, HCI reduced the variation in failure-to-rescue rates between hospitals by 2.7% after accounting for patient comorbidities and hospital resources.[55]

Hyder et al. estimated the improvement in mortality that may be realized through a reduction in FTR using the Nationwide Inpatient Sample (NIS, 2007-2011). High-mortality hospitals had higher FTR rates than low-mortality hospitals (22.4% vs 20.2%, p = 0.0020). Using Monte Carlo models they estimated that reducing the FTR gap by nearly 75% (2.73%; 95% CI 2.61 to 2.87) would potentially result in a 50% absolute reduction in baseline mortality (5.22%) for five target subpopulations.[56]

Indicator specifications

Several variations of the FTR measure are currently available, including the NQF endorsed measure from Children's Hospital of Philadelphia (NQF 0352) and PSI 04 (NQF 0351). In a comparison study, Silber at al showed the split sample reliability of the broader Silber FTR was higher than PSI 04 (0.32 vs. 0.18 vs. 0.18).[21] Silber's FTR is more highly correlated with 30 day mortality rate than PSI 04 (0.83 vs. 0.43). However, these results are expected given the broad scope of the Silber measure to capture most in-hospital deaths. This study may also suggest that the AHRQ measure captures a different aspect of quality care than broad mortality measures and a reduction in reliability may be acceptable given the focus on nursing sensitive complications.[57] Horwitz assessed the validity of V2.1 Rev 1 of PSI 04 using chart review data from the University Hospital Consortium. [58] In this study, using chart data as the gold study, they determined that 13.6% of the complications did not occur (15.8% of surgical cases), while an additional 8.1% met an AHRQ exclusion criterion but were incorrectly included. Since this study, the PSI 04 algorithm has been continuously improved as a result of learnings from studies of the PSI algorithms and user feedback.

Needleman et al., using PSI 04 (v3.1), examined whether the accuracy of PSI 04 (v3.1) could be improved by testing three present-on-admission (POA)-based exclusion rules using California HCUP Data.[16] Although these exclusion rules improved

the C-statistic of the failure-to-rescue measure, they did not affect the strong association between PSI 04 and higher nurse staffing and a greater percentage of registered nurses. The mortality rate was 22% among patients with hospital-acquired complications compared to 13% for patients with POA complications. Patients with hospital-acquired complications also had longer lengths of stay than patients with POA complications (14.1 versus 8.4 days; risk-adjusted difference = 4.3 days). The authors conclude that "failure-to-rescue is found in this analysis to be a robust measure and generates similar results regardless of how it is parameterized."[16]

References:

1. Silber J, Rosenbaum P, Williams S, Ross R, Schwartz J. The relationship between choice of outcome measure and hospital rank in general surgical procedures: implications for quality assessment. *Int J Qual Health Care* 1997;9(3):193-200.
2. Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. *Med Care.* 1992;30(7):615-629.
3. Johnston MJ, Arora S, King D, et al. A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. *Surgery.* 2015;157(4):752-763.
4. Silber JH, Kennedy SK, Even-Shoshan O, et al. Anesthesiologist direction and patient outcomes. *Anesthesiology.* 2000;93(1):152-163.
5. Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-staffing levels and the quality of care in hospitals. *N Engl J Med.* 2002;346(22):1715-1722.
6. Rosen AK, Rivard P, Zhao S, et al. Evaluating the patient safety indicators: How well do they perform on Veterans Health Administration data? . *Med Care.* 2005;43(9):873-884.
7. Needleman J, Buerhaus PI, Mattke S, Stewart M, Zelevinsky K. *Nurse Staffing and Patient Outcomes in Hospitals.* Boston, MA: Health Resources Services Administration; February 28, 2001 2001. 230-99-0021.
8. Manojlovich M, Talsma A. Identifying nursing processes to reduce failure to rescue. *J Nurs Adm.* 2007;37(11):504-509.
9. Schmid A, Hoffman L, Happ MB, Wolf GA, DeVita M. Failure to rescue: a literature review. *J Nurs Adm* 2007;37(4):188-198.
10. Silber JH, Rosenbaum PR, Ross RN. Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics? . *Journal of the American Statistical Association.* 1995;90(429):7-18.
11. Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA.* 2002;288(16):1987-1993.
12. Aiken LH, Clarke SP, Cheung RB, Sloane DM, Silber JH. Educational levels of hospital nurses and surgical patient mortality. *JAMA.* 2003;290(12):1617-1623.
13. Kendall-Gallagher D, Aiken LH, Sloane DM, Cimiotti JP. Nurse Specialty Certification, Inpatient Mortality, and Failure to Rescue. *J Nurs Scholarship.* 2011;43(2):188-194.
14. Friese CR, Aiken LH. Failure to rescue in the surgical oncology population: Implications for nursing and quality improvement. *Oncol Nurs Forum.* 2008;35(5):779-785.
15. Ghaferi AA, Osborne NH, Birkmeyer JD, Dimick JB. Hospital Characteristics Associated with Failure to Rescue from Complications after Pancreatectomy. *J Am Coll Surgeons.* 2010;211(3):325-330.
16. Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. *Med Care.* 2013;51(8):722-730.
17. Seago JA, Williamson A, Atwood C. Longitudinal analyses of nurse staffing and patient outcomes - More about failure to rescue. *J Nurs Admin.* 2006;36(1):13-21.
18. Boyle SM. Nursing unit characteristics and patient outcomes. *Nurs Econ.* 2004;22(3):111-+.
19. Clarke SP, Aiken LH. Failure to rescue. *Am J Nurs.* 2003;103(1):42-47.
20. Silber JH, Kennedy SK, Even-Shoshan O, et al. Anesthesiologist board certification and patient outcomes. *Anesthesiology.* 2002;96(5):1044-1052.
21. Silber JH, Romano PS, Rosen AK, Wang Y, Even-Shoshan O, Volpp KG. Failure-to-rescue: comparing definitions to measure quality of care. *Med Care.* 2007;45(10):918-925.
22. Blegen MA, Goode CJ, Park SH, Vaughn T, Spetz J. Baccalaureate education in nursing and patient outcomes. *J Nurs Adm.* 2013;43(2):89-94.
23. Neff DF, Cimiotti J, Sloane DM, Aiken LH. Utilization of non-US educated nurses in US hospitals: implications for hospital mortality. *Int J Qual Health Care.* 2013;25(4):366-372.
24. Park SH, Blegen MA, Spetz J, Chapman SA, De Groot H. Patient turnover and the relationship between nurse staffing and patient outcomes. *Res Nurs Health.* 2012;35(3):277-288.
25. Unruh LY, Zhang NJ. Nurse staffing and patient safety in hospitals: new variable and longitudinal approaches. *Nurs Res.* 2012;61(1):3-12.
26. Mills AC, Gillespie KN. Effect of Magnet hospital recognition on 2 patient outcomes. *J Nurs Care Qual.* 2013;28(1):17-23.

27. McHugh MD, Kelly LA, Smith HL, Wu ES, Vanak JM, Aiken LH. Lower mortality in magnet hospitals. *Med Care.* 2013;51(5):382-388.
28. Kutney-Lee A, Stimpfel AW, Sloane DM, Cimiotti JP, Quinn LW, Aiken LH. Changes in patient and nurse outcomes associated with magnet hospital recognition. *Med Care.* 2015;53(6):550-557.
29. Friese CR, Xia R, Ghaferi A, Birkmeyer JD, Banerjee M. Hospitals In 'Magnet' Program Show Better Patient Outcomes On Mortality Measures Compared To Non-'Magnet' Hospitals. *Health Affair.* 2015;34(6):986-992.
30. Mark BA, Harless DW, Spetz J, Reiter KL, Pink GH. California's minimum nurse staffing legislation: results from a natural experiment. *Health Serv Res.* 2013;48(2 Pt 1):435-454.
31. Ghaferi AA, Birkmeyer JD, Dimick JB. Hospital volume and failure to rescue with high-risk surgery. *Med Care.* 2011;49(12):1076-1081.
32. Mell MW, Kind A, Bartels CM, Smith MA. Failure to rescue and mortality after reoperation for abdominal aortic aneurysm repair. *J Vasc Surg.* 2011;54(2):346-351; discussion 351-342.
33. Gonzalez AA, Dimick JB, Birkmeyer JD, Ghaferi AA. Understanding the volume-outcome effect in cardiovascular surgery: the role of failure to rescue. *JAMA Surg.* 2014;149(2):119-123.
34. Castleberry AW, Clary BM, Migaly J, et al. Resident education in the era of patient safety: a nationwide analysis of outcomes and complications in resident-assisted oncologic surgery. *Ann Surg Oncol.* 2013;20(12):3715-3724.
35. Ferraris VA, Bolanos M, Martin JT, Mahan A, Saha SP. Identification of patients with postoperative complications who are at risk for failure to rescue. *JAMA Surg.* 2014;149(11):1103-1108.
36. Gopaldas RR, Overbey DM, Dao TK, Markley JG. The impact of academic calendar cycle on coronary artery bypass outcomes: a comparison of teaching and non-teaching hospitals. *J Cardiothorac Surg.* 2013;8:191.
37. Navathe AS, Silber JH, Small DS, et al. Teaching hospital financial status and patient outcomes following ACGME duty hour reform. *Health Serv Res.* 2013;48(2 Pt 1):476-498.
38. Ghaferi AA, Birkmeyer JD, Dimick JB. Complications, failure to rescue, and mortality with major inpatient surgery in medicare patients. *Ann Surg.* 2009;250(6):1029-1034.
39. Wakeam E, Hyder JA, Tsai TC, Lipsitz SR, Orgill DP, Finlayson SR. Complication timing and association with mortality in the American College of Surgeons' National Surgical Quality Improvement Program database. *J Surg Res.* 2015;193(1):77-87.
40. Silber JH, Rosenbaum PR, Kelz RR, et al. Examining Causes of Racial Disparities in General Surgical Mortality: Hospital Quality Versus Patient Risk. *Med Care.* 2015;53(7):619-629.
41. Carthon JM, Kutney-Lee A, Jarrin O, Sloane D, Aiken LH. Nurse staffing and postsurgical outcomes in black adults. *J Am Geriatr Soc.* 2012;60(6):1078-1084.
42. Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. *JAMA Surg.* 2014;149(5):475-481.
43. Schneider EB, Ejaz A, Spolverato G, et al. Hospital volume and patient outcomes in hepato-pancreatico-biliary surgery: is assessing differences in mortality enough? *J Gastrointest Surg.* 2014;18(12):2105-2115.
44. Spolverato G, Ejaz A, Hyder O, Kim Y, Pawlik TM. Failure to rescue as a source of variation in hospital mortality after hepatic surgery. *Br J Surg.* 2014;101(7):836-846.
45. Wright JD, Herzog TJ, Siddiq Z, et al. Failure to rescue as a source of variation in hospital mortality for ovarian cancer. *J Clin Oncol.* 2012;30(32):3976-3982.
46. Yasunaga H, Hashimoto H, Horiguchi H, Miyata H, Matsuda S. Variation in cancer surgical outcomes associated with physician and nurse staffing: a retrospective observational study using the Japanese Diagnosis Procedure Combination Database. *BMC Health Serv Res.* 2012;12:129.
47. Friese CR, Earle CC, Silber JH, Aiken LH. Hospital characteristics, clinical severity, and outcomes for surgical oncology patients. *Surgery.* 2010;147(5):602-609.
48. LaPar DJ, Ghanta RK, Kern JA, et al. Hospital variation in mortality from cardiac arrest after cardiac surgery: an opportunity for improvement? *Ann Thorac Surg.* 2014;98(2):534-539; discussion 539-540.
49. Wakeam E, Hevelone ND, Maine R, et al. Failure to rescue in safety-net hospitals: availability of hospital resources and differences in performance. *JAMA Surg.* 2014;149(3):229-235.
50. Sacks GD, Lawson EH, Dawes AJ, et al. Relationship Between Hospital Performance on a Patient Satisfaction Survey and Surgical Quality. *JAMA Surg.* 2015;150(9):858-864.
51. Brooke BS, Dominici F, Pronovost PJ, Makary MA, Schneider E, Pawlik TM. Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery.* 2012;151(5):651-659.
52. Young-Xu YN, Fore AM, Metcalf A, Payne K, Neily J, Sculli GL. Using Crew Resource Management and a 'Read-and-Do Checklist' to Reduce Failure-to-Rescue Events on a Step-Down Unit. *Am J Nurs.* 2013;113(9):51-57.
53. Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. *N Engl J Med.* 2009;361(14):1368-1375.

54. Friese C, Lake E, Aiken LH, Silber JH, Sochalski J. Hospital nurse practice environments and outcomes for surgical oncology patients. *Health Serv Res.* 2008;43(4):1145-1163.
55. Sheetz KH, Dimick JB, Ghaferi AA. The association between hospital care intensity and surgical outcomes in medicare patients. *JAMA Surg.* 2014;149(12):1254-1259.
56. Hyder JA, Wakeam E, Adler JT, DeBord Smith A, Lipsitz SR, Nguyen LL. Comparing Preoperative Targets to Failure-to-Rescue for Surgical Mortality Improvement. *J Am Coll Surg.* 2015;220(6):1096-1106.
57. Needleman J, Buerhaus PI. Failure-to-rescue: Comparing definitions to measure quality of care. *Medical Care.* 2007;45(10):913-915.
58. Horwitz LI, Cuny JF, Cerese J, Krumholz HM. Failure to rescue - Validation of an algorithm using administrative data. *Medical Care.* 2007;45(4):283-287.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PSI04_Measure_Evidence_Form_160531_v2.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
This indicator targets patients who are admitted for surgery who die following the development of a serious but treatable complication of care. Examples of such complications include: 1) shock or cardiac arrest, 2) sepsis, 3) pneumonia, 4) deep vein thrombosis or pulmonary embolism, and 5) gastrointestinal hemorrhage or acute ulcer. This indicator is fundamentally different than other PSIs, as it reflects the effectiveness of the hospital in rescuing a patient from complications versus preventing the underlying complications.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
These tables (1a. and 1b.) are also included in the supplemental files.

Table 1a. Reference Population Observed Rate for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04), 2011-2013

Overall Reference Population Rate

Year2	Number of Hospitals (Numerator)1	Outcome of Interest Population at Risk (Denominator)1	Observed Rate Per 1000 Surgical Discharges1
2013	2,783	21,242	182,512 116.3869
2012	2,860	21,897	185,872 117.8069
2011	2,748	21,403	181,317 118.0419

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)
1The observed rate refers to the total rate for all observations included in the reference population data (numerator) divided by the total combined eligible population of all hospitals included in the reference population data (denominator).

2Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

Table 1b. Distribution of Hospital Performance for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) in 2-year Pooled Data (2011-2012, 2012-2013)1

Distribution of Hospital-level Observed Rates in Reference Population

Year3	Number of Hospitals	Rates per 1000 Surgical Discharges (p=percentile)2							
		Mean	SD2	p5	p25	Median	p75	p95	
2011-2012	3,212	103.83	85.55	0.00	58.06	102.61	140.35	217.39	
2012-2013	3,398	100.76	88.28	0.00	51.28	99.92	137.25	212.12	

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

1Consistent with the recommended minimum reporting time period, results are presented for data combining 2 years of data: 2011 and 2012, 2012 and 2013. Data from 2012 are included in both time periods reported. Limitations in present on admission data (POA) data availability (see below) do not allow for use of earlier years.

2The distribution of hospital rates reports the mean and standard deviation (SD) of the observed rates for all hospitals in the dataset with at least one case in the denominator, as well as the observed rate for hospitals in the 5th, 25th, 50th (median), 75th, and 95th percentile. Standard deviation refers to the spread in observed values in relation to the mean.

3Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

Not applicable

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use*.

Table 2 shows that the risk of death for surgical inpatients with serious treatable complications varies by age, sex, community-level income, expected payer, and region. In 2013, among 182,512 surgical patients with serious treatable conditions, older patients, men, those from lower income communities, those with Medicare, Medicaid or uninsured, and those treated in the Northeast were at greater risk of death (after controlling for a variety of clinical risk factors) than younger patients, women, those from higher income communities, those with private insurance and those treated in the Midwest or West. These findings are based on 182,512 discharges from 2,783 hospitals in the 34 states in 2013 and reflect national population estimates. The findings may be different at an individual hospital-level.

Please note: Table 2, as shown below, is unformatted and may be difficult to read. A formatted Table 2 is provided in the supplemental files.

Table 2. Risk-Adjusted Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) per 1,000 surgical discharges, by patient and hospital characteristics, 2013

Patient/hospital characteristic	Risk-adjusted Estimate1	Std Errorp-value	(Ref Grp = *)	Lower 95% CL	Upper 95% CL
Total U.S.	116.387	0.675		115.063	117.711
Patient Characteristics					
Age Groups:					
18-39	74.420	2.443	*	69.631	79.209
40-64	100.445	1.105	<.001	98.280	102.611
65 and over	133.431	0.923	<.001	131.621	135.241
Gender:					
Male2	120.100	0.909	*	118.319	121.882
Female	111.841	1.010	<.001	109.860	113.821
Patient Zip Code Median Income					

First quartile (lowest income)			121.085	2.383	<.001	116.415	125.755
Second quartile	118.659	1.532	<.001	115.656	121.661		
Third quartile	120.176	1.393	<.001	117.445	122.907		
Fourth quartile (highest income)2			112.644	0.969	*	110.746	114.543
Location of patient residence (NCHS)3:							
Rural	121.536	5.067	0.148	111.605	131.468		
Urban2	116.192	0.683	*	114.854	117.531		
Expected payment source:							
Private insurance2			109.914	1.534	*	106.906	112.921
Medicare	116.361	0.838	<.001	114.718	118.004		
Medicaid	120.407	2.300	<.001	115.899	124.915		
Uninsured / self-pay / no charge			136.964	3.203	<.001	130.687	143.241
Other insurance	116.814	4.217	0.062	108.549	125.078		
Location of Care:							
Northeast2			122.821	1.769	*	119.355	126.288
Midwest	110.965	1.440	<.001	108.143	113.788		
South	118.210	1.097	0.013	116.061	120.359		
West	115.587	1.375	<.001	112.893	118.282		

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

1Rates are adjusted using the AHRQ QI PSI POA Reference Population for 2013 as the standard population. Age and gender are removed from models for the relevant strata.

2Reference group

3NCHS - National Center for Health Statistics designation for urban-rural locations. Metropolitan areas are considered urban and micropolitan or non-core areas are considered rural.

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.

Not applicable

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Patient/societal consequences of poor quality

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

This indicator was originally proposed by Silber et al. as a more powerful tool than the risk-adjusted mortality rate to detect true differences in patient outcomes across hospitals.1 The underlying premise was that better hospitals are distinguished not by having fewer adverse occurrences but by more successfully averting death among (i.e., rescuing) patients who experience such complications. Silber et al's original definition was based on key clinical findings abstracted from the medical records of 2,831 cholecystectomy patients and 3,141 transurethral prostatectomy patients admitted to 531 hospitals in 1985.1 The key postoperative diagnoses that defined the denominator at risk of "failure to rescue" (FTR) included cardiac arrhythmias, congestive heart failure, cardiac arrest, pneumonia, pulmonary embolus, pneumothorax, renal dysfunction, stroke, wound infection, and unplanned return to surgery.1 More recently, Needleman and Buerhaus adapted failure to rescue to administrative data sets, with a specific focus on optimizing the sensitivity of the indicator to nurse staffing and skill mix. Their denominator definition included the ICD-9-CM codes for sepsis, pneumonia (including aspiration), acute upper gastrointestinal bleeding, shock, cardiac/respiratory arrest, deep vein thrombosis (DVT), and pulmonary embolus (PE).2 Both specifications have been linked to factors such as board-certified anesthesiologists, board certified surgeons, registered nurse staffing ratios, nursing skill mix, and other structural and processes measures of high-quality care. 1-18 Due to improvements in care, between 1998 and 2007, FTR decreased by 6.05% per year

(p<0.0001).19 However, Table 1 above shows that PSI 04 still captures approximately 43,000 deaths each year in the 34 states in the all-payer reference population.

1c.4. Citations for data demonstrating high priority provided in 1a.3

1. Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. Med Care. 1992;30(7):615-629.
2. Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-staffing levels and the quality of care in hospitals. N Engl J Med. 2002;346(22):1715-1722.
3. Navathe AS, Silber JH, Small DS, et al. Teaching hospital financial status and patient outcomes following ACGME duty hour reform. Health Serv Res. 2013;48(2 Pt 1):476-498.
4. Needleman J, Buerhaus PI, Mattke S, Stewart M, Zelevinsky K. Nurse Staffing and Patient Outcomes in Hospitals. Boston, MA: Health Resources Services Administration; February 28, 2001 2001. 230-99-0021.
5. Manojlovich M, Talsma A. Identifying nursing processes to reduce failure to rescue. J Nurs Adm. 2007;37(11):504-509.
6. Schmid A, Hoffman L, Happ MB, Wolf GA, DeVita M. Failure to rescue: a literature review. J Nurs Adm 2007;37(4):188-198.
7. Silber JH, Rosenbaum PR, Ross RN. Comparing the contributions of groups of predictors: Which outcomes vary with hospital rather than patient characteristics? . Journal of the American Statistical Association. 1995;90(429):7-18.
8. Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. Jama-J Am Med Assoc. 2002;288(16):1987-1993.
9. Aiken LH, Clarke SP, Cheung RB, Sloane DM, Silber JH. Educational levels of hospital nurses and surgical patient mortality. JAMA. 2003;290(12):1617-1623.
10. Kendall-Gallagher D, Aiken LH, Sloane DM, Cimiotti JP. Nurse Specialty Certification, Inpatient Mortality, and Failure to Rescue. J Nurs Scholarship. 2011;43(2):188-194.
11. Friese CR, Aiken LH. Failure to rescue in the surgical oncology population: Implications for nursing and quality improvement. Oncol Nurs Forum. 2008;35(5):779-785.
12. Ghaferi AA, Osborne NH, Birkmeyer JD, Dimick JB. Hospital Characteristics Associated with Failure to Rescue from Complications after Pancreatectomy. J Am Coll Surgeons. 2010;211(3):325-330.
13. Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. Med Care. 2013;51(8):722-730.
14. Seago JA, Williamson A, Atwood C. Longitudinal analyses of nurse staffing and patient outcomes - More about failure to rescue. J Nurs Admin. 2006;36(1):13-21.
15. Boyle SM. Nursing unit characteristics and patient outcomes. Nurs Econ. 2004;22(3):111-+.
16. Clarke SP, Aiken LH. Failure to rescue. Am J Nurs. 2003;103(1):42-47.
17. Silber JH, Kennedy SK, Even-Shoshan O, et al. Anesthesiologist direction and patient outcomes. Anesthesiology. 2000;93(1):152-163.
18. Johnston MJ, Arora S, King D, et al. A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. Surgery. 2015;157(4):752-763.
19. Downey JR, Hernandez-Boussard T, Banka G, Morton JM. Is patient safety improving? National trends in patient safety indicators: 1998-2007. Health Serv Res. 2012;47(1 Pt 2):414-430.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*

Not applicable

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Complications

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://1.usa.gov/1TksC2k Note: The URL link will be updated for version 6.0 public release found via the module page: http://qualityindicators.ahrq.gov/Modules/psi_resources.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PSI04_Technical_Specifications_v6.0_160527.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
As standard protocol, the AHRQ QI program annually updates all measures with Fiscal Year coding changes, refinements based on stakeholder input, refinements to improve specificity and sensitivity based on additional analyses, and necessary software changes. In addition, approximately every two years, AHRQ updates the risk-adjustment parameter estimates based on the most recent year of data (i.e., the most current reference population possible). The refined measures are tested and confirmed to be valid and reliable prior to release of the updated software.

Since the last update, the following changes have been made to the indicator:
This revised version will be implemented in forthcoming version 6.0 specifications in 2016. This version (v6.0) includes the following changes from the previously-endorsed version (v4.4):

- An ICD-10-CM/PCS version has been created.
- STRATUM_SHOCK (previously Stratum D):
  - Abortion-related shock diagnosis codes added to Denominator
    63450 - SPON ABORT W SHOCK-UNSP
    63451 - SPON ABORT W SHOCK-INC
    63452 - SPON ABORT W SHOCK-COMP
    63550 - LEGAL ABORT W SHOCK-UNSO
    63551 - LEGAL ABORT W SHOCK-INC
    63552 - LEGAL ABORT W SHOCK-COMP
    63650 - ILLEG AB W SHOCK-UNSO
    63651 - ILLEG ABORT W SHOCK-INC
    63652 - ILLEG ABORT W SHOCK-COMP
    63750 - ABORT NOS W SHOCK-UNSO
    63751 - ABORT NOS W SHOCK-INC
    63752 - ABORT NOS W SHOCK-COMP
    6385 - ATTEM ABORTION W SHOCK
  - Codes removed from denominator (eliminating overlap with STRATUM_SEPSIS):
    78552 – SEPTIC SHOCK
    99802 – POSTOP SHOCK, SEPTIC
  - Code added to Denominator principal diagnosis exclusion: 53021 (ULCER OF ESOPHAGUS WITH BLEEDING)
- STRATUM_SEPSIS (previously Stratum C):
  - Codes removed from Denominator (eliminating overlap with STRATUM_SHOCK):
    78559 (SHOCK W/O TRAUMA NEC)
    99800 (POSTOPERATIVE SHOCK, NOS)
  - Codes added to the Denominator Exclusion:

70700 PRESSURE ULCER, SITE NOS
70701 PRESSURE ULCER, ELBOW
70702 PRESSURE ULCER, UPR BACK
70703 PRESSURE ULCER, LOW BACK
70704 PRESSURE ULCER, HIP
70705 PRESSURE ULCER, BUTTOCK
70706 PRESSURE ULCER, ANKLE
70707 PRESSURE ULCER, HEEL
70709 PRESSURE ULCER, SITE NEC

- STRATUM_PNEUMONIA (previously Stratum B):
  - Codes added to Denominator: 481 – PNEUMOCOCCAL PNEUMONIA [STREPTOCOCCUS PNEUMONIAE PNEUMONIA
  - Codes added to Denominator exclusions: ICD-9-CM Lung cancer procedure codes for thoracoscopic surgery (3230, 3241, 3250)
- STRATUM_DVT (previously Stratum A):
  - Codes removed from Denominator: 45342 (AC DVT/EMB DISTL LOW EXT),
- STRATUM_GI_HEM (previously Stratum E)

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Number of deaths (DISP=20) among cases meeting the inclusion and exclusion rules for the denominator.

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
For users with a complete all-payer sample of hospital discharge, the recommended time period is two years for measurement of hospital rates. This recommendation is based on testing of reliability of the measure; this reliability testing is specific to all-payer hospital populations. Reliability estimates often vary when the measure is applied to other hospital populations, such as Medicare-only populations. Reliability is sensitive to numerator and denominator size as well as the distribution of hospital rates. For populations other than all-payer hospital populations fewer or more than 2 years of data may be recommended, depending on changes in reliability estimates. Note that the signal variance parameters embedded in the AHRQ QI software assume at least a one-year time period. Users may use longer time periods if desired.

S.6. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Please see attached excel file in S.2b. for version 6.0 specifications.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Surgical discharges, for patients ages 18 through 89 years or MDC 14 (pregnancy, childbirth, and puerperium), with all of the following:

- any-listed ICD-9-CM or ICD-10-PCS procedure codes for an operating room procedure; and
- the principal procedure occurring within 2 days of admission or an admission type of elective (ATYPE=3); and
- meet the inclusion and exclusion criteria for STRATUM_SHOCK (shock or cardiac arrest), STRATUM_SEPSIS (sepsis), STRATUM_PNEUMONIA (pneumonia), STRATUM_DVT (deep vein thrombosis or pulmonary embolism), or STRATUM_GI_HEM (gastrointestinal hemorrhage or acute ulcer)

STRATUM_SHOCK (shock or cardiac arrest)

- any secondary ICD-9-CM or ICD-10-CM diagnosis codes or any-listed ICD-9-CM or ICD-10-PCS procedure codes for shock or cardiac arrest

STRATUM_SEPSIS (sepsis)

- any secondary ICD-9-CM or ICD-10-CM diagnosis codes for sepsis.

STRATUM_PNEUMONIA (pneumonia)

- any secondary ICD-9-CM or ICD-10-CM diagnosis codes for pneumonia or pneumonitis.

STRATUM_DVT (deep vein thrombosis or pulmonary embolism)
- any secondary ICD-9-CM or ICD-10-CM diagnosis codes for deep vein thrombosis or pulmonary embolism.

STRATUM_GI_HEM (gastrointestinal hemorrhage or acute ulcer)
- any secondary ICD-9-CM or ICD-10-CM diagnosis codes for gastrointestinal hemorrhage or acute ulcer.

Surgical discharges are defined by specific MS-DRG codes and ICD-9-CM/ICD-10-PCS codes indicating "major operating room procedures."

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Please see attached excel file in S.2b. for v6.0 specifications.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Exclude cases:
- transferred to an acute care facility (DISP = 2)
- with missing discharge disposition (DISP=missing), gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing)

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Please see attached excel file in S.2b. for v6.0 specifications.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
Please see attached excel file in S.2b. for v6.0 specifications.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
The predicted value for each case is computed using a hierarchical model (logistic regression with hospital random effect) and covariates for gender, age (in 5-year age groups, except for the youngest age range), Modified Diagnosis Related Groups (ie. MS-DRGs without any distinction for "comorbidity and complications" (CC/MCC), Elixhauser Comorbidity Index (https://www.hcup-us.ahrq.gov/toolssoftware/comorbidity/comorbidity.jsp), Major Diagnosis Categories (MDC) based on the principal diagnosis, and transfer in from another acute care hospital. A parsimonious model was identified using a backward stepwise selection procedure with bootstrapping. The expected rate is computed as the sum of the predicted value for each case divided by the number of cases for the unit of analysis of interest (i.e., hospital). The risk-adjusted rate for the overall PSI 04 is calculated as the observed to expected ratio multiplied by the reference population rate, where the observed and expected values are summed across five strata (categories) of PSI 04 risk. This approach differs from other AHRQ Patient Safety Indicators without strata, in that each discharge-record's expected value is computed using one of five distinct stratum-specific risk adjustment models that correspond to an assigned PSI 04 stratum. The five PSI 04 strata group records together based on secondary diagnoses that represent complications of care, and place the patient at risk of death (which is the numerator of PSI 04).

Additional information on methodology can be found in the Empirical Methods document on the AHRQ Quality Indicator website (www.qualityindicators.ahrq.gov). The Empirical Methods are also attached in the supplemental materials.

The specific covariates for this measure are provided for each Stratum as part of the Technical Specifications attached to section S.2b.
Source: http://www.qualityindicators.ahrq.gov/Modules/psi_resources.aspx

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not applicable.

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The observed rate is the number of discharge records where the patient experienced the PSI adverse event divided by the number of discharge records at risk for the event. The expected rate is a comparative rate that incorporates information about a reference population that is not part of the user's input dataset – what rate would be observed if the expected level of care observed in the reference population and estimated with risk adjustment regression models, were applied to the mix of patients with demographic and comorbidity distributions observed in the user's dataset. The expected rate is calculated only for risk-adjusted indicators.

The following descriptions are for the expected rate and risk-adjusted rate. These rates are calculated using models for each individual stratum.

The expected rate is estimated using the stratum specific model for each record using a generalized estimating equations (GEE) approach to account for correlation at the hospital or provider level. Records are assigned to the stratum for which they qualify with the highest observed mortality rate.

The risk-adjusted rate is a comparative rate that also incorporates information about a reference population that is not part of the input dataset – what rate would be observed if the level of care observed in the user's dataset were applied to a mix of patients with demographics and comorbidities distributed like the reference population? The risk-adjusted rate for the overall PSI 04 is calculated as the observed to expected ratio multiplied by the reference population rate, where the observed and expected values are summed across five strata (categories) of PSI 04 risk. This approach differs from other AHRQ Patient Safety Indicators without strata, in that each discharge-record's expected value is computed using one of five distinct stratum-specific risk adjustment models that correspond to an assigned PSI 04 stratum. The five PSI 04 strata group records together based on secondary diagnoses that represent complications of care, and place the patient at risk of death (which is the numerator of PSI 04).

The smoothed rate is the weighted average of the risk-adjusted rate from the user's input dataset and the rate observed in the reference population; the smoothed rate is calculated with a shrinkage estimator to result in a rate near that from the user's dataset if the provider's rate is estimated in a stable fashion with minimal noise, or to result in a rate near that of the reference population if the variance of the estimated rate from the input dataset is large compared with the hospital-to-hospital variance estimated from the reference population. Thus, the smoothed rate is a weighted average of the risk-adjusted rate and the reference population rate, where the weight is the signal-to-noise ratio. In practice, the smoothed rate brings rates toward the mean, and tends to do this more so for outliers (such as rural hospitals).

For additional information, please see the supplemental materials for the AHRQ QI Empirical Methods.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not applicable

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not applicable

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Exclude cases with missing gender (SEX=missing), age (AGE=missing), quarter (DQTR=missing), year (YEAR=missing), or principal diagnosis (DX1=missing). Missingness on these variables, in aggregate, almost never exceeds 1% of eligible records.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
While the measure is tested and specified using data from the Healthcare Cost and Utilization Project (HCUP) (see section 1.1 and 1.2 of the measure testing form), the measure specifications for numerators, denominators and observed rates and software are specified to be used with any ICD-9-CM- or ICD-10-CM/PCS coded administrative billing/claims/discharge dataset. Software to calculate risk-adjusted and smoothed rates is available for ICD-9-CM only. One year of ICD-10-CM/PCS coded data is necessary before risk adjustment will be available for ICD-10-CM/PCS versions of the software.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not applicable

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PSI04_Measure_Testing_Form_160615.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0351
Measure Title: Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04)
Date of Submission: 5/31/2016

Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

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). [13]

2b4. 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 sociodemographic 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.

2b5. 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.
2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.
2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes
10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

All analyses were completed using data from the Healthcare Cost and Utilization Project (HCUP) State Inpatient Databases (SID), 2011-2013. HCUP is a family of health care databases and related software tools and products developed through a Federal-State-Industry partnership and sponsored by the Agency for Healthcare Research and Quality (AHRQ).[1] HCUP databases bring together the data collection efforts of State data organizations, hospital associations, private data organizations, and the Federal government to create a national information resource of encounter-level health care data. The HCUP SID contain the universe of the inpatient discharge abstracts in participating States, translated into a uniform format to facilitate multi-State comparisons and analyses. All states provide data for community hospitals and together, the SID encompasses about 97 percent of all U.S. community hospital discharges. For the analyses presented here, we use 34 states representing about 89 percent of the U.S. community hospital discharges, for a total of about 30 million hospital discharges from community hospitals. As defined by the American Hospital Association, community hospitals are all non-Federal, short-term, general or other specialty hospitals, excluding hospital units of institutions. Included among community hospitals are public and academic medical centers, specialty hospitals such as obstetrics–gynecology, ear–nose–throat, orthopedic and pediatric institutions. Short-stay rehabilitation, long-term acute care hospitals are excluded from the data used for the reported analyses.

Each of the 34 states included in the dataset report information about whether a diagnosis was present on admission (POA) and information on the timing of procedures during the hospitalization. POA data[2] is important to distinguish complications that occur in-hospital from diagnoses that existed prior to hospitalization. Edit checks on POA were developed using a separate analysis of HCUP databases that examined POA coding in the 2013 SID at hospitals that were required to report POA to CMS. The edits identify general patterns of suspect reporting of POA. The edits do not evaluate whether a valid POA value (e.g., Y or N) is appropriate for the specific diagnosis. There are three hospital-level edit checks:

1. Indication that a hospital has POA reported as Y on all diagnoses on all discharges
2. Indication that a hospital has POA reported as missing on all non-Medicare discharges
3. Indication that a hospital reported POA as missing on all nonexempt diagnoses for 15 percent or more of discharges. The cut-point of 15 percent was determined by 2 times the standard deviation plus the mean of the percentage for hospitals required to report POA to CMS.

Hospitals that failed any of the edit checks were excluded from the dataset.
The SID data elements include International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) coded principal and secondary diagnoses and procedures, additional detailed clinical and service information based on revenue codes, admission source and discharge status, patient demographics, expected payment source (Medicare, Medicaid, private insurance as well as the uninsured), total charges and length of stay (www.hcup-us.ahrq.gov).

1.3. What are the dates of the data used in testing?

HCUP data included calendar years 2011-2013. Further explanation of the years used for each analysis are in section 1.7.

1.4. What levels of analysis were tested? (*testing must be provided for <u>all</u> the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency

[1] HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011-2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. (AHRQ QI Software Version 6.0)
[2] Present-on -Admission was added as a data element to the uniform bill form (UB-04) effective October 1, 2007, and hospitals incurred a payment penalty for not including POA on Medicare records beginning October 1, 2008. Each of the several diagnoses in a discharge record can be flagged as "present at the time the order for inpatient admission occurs" or not (see http://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm).

☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which <u>measured entities</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Table 1a. Reference Population Observed Rate for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04), 2011-2013

Overall Reference Population Rate				
Year[2]	Number of Hospitals	Outcome of Interest (Numerator)[1]	Population at Risk (Denominator)[1]	Observed Rate Per 1000 Surgical Discharges[1]
2013	2,783	21,242	182,512	116.3869
2012	2,860	21,897	185,872	117.8069
2011	2,748	21,403	181,317	118.0419

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

[1]The observed rate refers to the total rate for all observations included in the reference population data (numerator) divided by the total combined eligible population of all hospitals included in the reference population data (denominator).

[2]Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

Table 1b. Distribution of Hospital Performance for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) in 2-year Pooled Data (2011-2012, 2012-2013)[1]

Distribution of Hospital-level Observed Rates in Reference Population								
Year[3]	Number of Hospitals	Rates per 1000 Surgical Discharges (p=percentile)[2]						
		Mean	SD[2]	p5	p25	Median	p75	p95
2011-2012	3,212	103.83	85.55	0.00	58.06	102.61	140.35	217.39
2012-2013	3,398	100.76	88.28	0.00	51.28	99.92	137.25	212.12

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2011 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

[1]Consistent with the recommended minimum reporting time period, results are presented for data combining 2 years of data: 2011 and 2012, 2012 and 2013. All data from 2012 are included in both time periods reported. Limitations in present on admission data (POA) data availability (see below) do not allow for use of earlier years.

[2]The distribution of hospital rates reports the mean and standard deviation (SD) of the observed rates for all hospitals in the dataset with at least one case in the denominator, as well as the observed rate for hospitals in the 5th, 25th, 50th (median), 75th, and 95th percentile. Standard deviation refers to the spread in observed values in relation to the mean.

[3]Reference population is limited to states with present on admission data (POA). Since many states did not report POA data prior to 2011 we have not included testing prior to 2011.

1.6. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

See 1.5 (Table 1a)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

For tests requiring hospital rates we combine 2 years of hospital data prior to calculating rates and testing the measure (termed "2-year pooled data"). The tests that used pooled 2012 and 2013 data include: reliability testing (Table 2), validity testing (described in text, section 2.b) and performance discrimination (Table 7). The hospital rate distributions (Table 1b) are reported for two 2-year pooled data periods, 2011 – 2012 and 2012 – 2013. All other tests that do not use hospital rates are calculated using 2013 data.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

Age and sex were the only patient sociodemographic characteristics that were available and analyzed in the data used for measure development and testing. Many of the HCUP SID include race/ethnicity, and all of the HCUP SID include the primary expected source of payment and zip code of residence, which could be used to capture socioeconomic characteristics at an ecological (community) level. While these variables were used to assess disparities at the national level, these variables were not used in the current risk adjustment model, based on our conceptual description (i.e., logical rationale or theory informed by literature and content experts) of the causal pathway between these factors, patient clinical factors, quality of care, and outcome, described in Section 2b4.3 below.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

The AHRQ QIs use signal-to-noise ratios to assess reliability**.** The signal-to-noise ratio is a measure of reliability that is calculated at the hospital level and then summarized across the entire population of US hospitals. It compares the degree to which risk adjusted rates differ across hospitals (the signal) to the degree of precision of the rates within hospitals (the noise). This metric is a stringent measure of reliability that takes into account the observed distribution of risk adjusted rates within a reference population. An indicator with a low signal-to-noise ratio may not be able to distinguish differences in performance among hospitals, or may identify differences inconsistently within the same time period. An indicator with a high signal-to-noise ratio will be more likely to consistently distinguish performance differences among hospitals (e.g. one hospital performs better than others).

The signal-to-noise ratio is estimated for each hospital. The overall signal-to-noise estimate is an average of hospital-level signal to noise ratios weighted by a value of one divided by the signal plus the hospital's noise for PSI 04. Hospitals with smaller denominators (the number of patients at risk) will have lower weight, and less influence on the overall signal-to-noise ratio, because of higher noise. Weighting reduces the influence of hospitals that have less reliable rates due to very small denominators (the number of patients at risk) on the overall signal-to-noise ratio estimate.

Because the signal-to-noise ratio quantifies the ability to consistently discriminate one hospital's performance from the other hospitals in the population, it is sensitive to the distribution of hospital sizes as well as the distribution of risk-adjusted rates in the reference population. If the hospitals in a population all have performance in a narrow range (low signal), it is more difficult to reliably distinguish among hospitals' performance than when hospital performance is spread out over a much wider range (high signal). For example, if all hospitals have nearly perfect performance, it will be impossible to distinguish among them. As a consequence, if the distribution of hospital rates changes over time, or if the measured population is restricted (e.g. Medicare patients), or if a different subset of hospitals is included, the signal-to-noise ratio will also change.

There is no universally accepted threshold of "adequate" signal to noise ratio. Different methods of calculating reliability and signal-to-noise (e.g., split sample or test-retest reliability of the data, different methods of calculating the hospital signal-to-noise ratio) result in different distributions of reliability scores. In addition, "adequate" depends on the specific application and judgment of the user. For instance, if a complication such as mortality is very important (e.g. leads to great harm to the patient) a lower reliability may be acceptable. However, the AHRQ QI program generally considers ratios between 0.4 – 0.8 as acceptable. It is rare to achieve reliability above 0.8, using hospital signal-to-noise ratios as an indicator of reliability. To account for the uncertainty (noise) in a hospital's performance due to low volume, a longer period of data can be used or smoothed rates can be calculated.

For reference, the following text in black was previously submitted to NQF. Most of the information is outdated. PSI 4 A higher risk-adjusted mortality rate for death among surgical inpatients with serious treatable complications is associated with significantly higher costs. The AHRQ QIs have the advantage of taking the multidimensional nature of hospital quality into account. As the coefficients on the AHRQ QIs show, measures of hospital quality can have conflicting effects on hospital costs. A single measure that combines these effects into one variable offers less insight into hospital performance than the outcomes for each measure. [1]

Patient Safety Events Are Common at U.S. Hospitals: Between 2005 and 2007 there were 913,215 total patient safety events among Medicare beneficiaries. Common Patient Safety Events are Very Costly: Between 2005 and 2007 these patient safety events were associated with over $6.9 billion of wasted healthcare cost. Less Improvement Seen Among Most Common Events: Eight patient safety indicators showed improvement while seven indicators worsened in 2007 compared to 2005. Some of the most common and most serious indicators worsened, including decubitus ulcer (bed sores), sepsis, respiratory failure, deep vein thrombosis (blood clots in the legs), and pulmonary embolism (potentially fatal blood clots forming in the lungs). Approximately One-in-Ten Medicare Patients with Patient Safety Events Died: Between 2005 and 2007 there were 97,755 actual inhospital deaths that occurred among patients who experienced one or more of the 15 patient safety events. [2]

PSI 4: death among surgical inpatients with serious treatable complications was not included because many procedure codes are required. [3]

The initial translation (electronic mapping, review and revision by expert coder, programming of codes and testing on data from 1996-1998 [ICD 9-CM] to 1998-2006 [ICD-10-AM, through 4 editions]) found that differences between ICD-9-CM and ICD-10-AM datasets presented some challenges. After this phase, which was faithful to AHRQ´s case definitions, the indicators were refined for use with the condition onset flag, resulting in the AusPSIs. [4]

Principal Findings. Excess 90-day expenditures likely attributable to PSIs ranged from $646 for technical problems (accidental laceration, pneumothorax, etc.) to $28,218 for acute respiratory failure, with up to 20 percent of these costs incurred postdischarge. With a third of all 90-day deaths occurring postdischarge, the excess death rate associated with PSIs ranged from 0 to 7 percent. The excess 90-day readmission rate associated with PSIs ranged from 0 to 8 percent. Overall, 11 percent of all deaths, 2 percent of readmissions, and 2 percent of expenditures were likely due to these 14 PSIs. Conclusions. The effects of medical errors continue long after the patient leaves the hospital. Medical error studies that focus only on the inpatient stay can underestimate the impact of patient safety events by up to 20-30 percent. [5]

AHRQ 2007 State Inpatient Databases (SID) with 4,000 hospitals and 30 million discharges

References

[1] Laditka JN, Laditka SB, Cornman CB. Evaluating hospital care for individuals with Alzheimer´s disease using inpatient quality indicators. Am J Alzheimers Dis Other Demen. 2005 Jan-Feb;20(1):27-36. PMID: 15751451.

[2] HealthGrades. Every 1.7 Minutes a Medicare Beneficiary Experiences a Patient Safety Event. Business Wire. Available on-line: http://www.allbusiness.com/government/government-bodies-offices/12279340-1.html. Accessed 1/11/2011.

[3] Hude Quan, MD, PhD; Saskia Drösler, MD; Vijaya Sundararajan, et al. Adaptation of AHRQ Patient Safety Indicators for Use in ICD-10 Administrative Data by an International Consortium. In Advances in Patient Safety: New Directions and Alternative Approaches (Vol. 1: Assessment). Henriksen K, Battles JB, Keyes MA, et al., editors. Rockville (MD): Agency for Healthcare Research and Quality; 2008 Aug. Bookshelf ID: NBK43634.

[4] McConchie S, Shepheard J, Waters S, McMillan AJ, Sundararajan V. The AusPSIs: the Australian version of the Agency of Healthcare Research and Quality patient safety indicators. Aust Health Rev. 2009 May;33(2):334-41. PMID: 19563325.

[5] Encinosa WE, Hellinger FJ. The impact of medical errors on ninety-day costs and outcomes: an examination of surgical patients. Health Serv Res. 2008 Dec;43(6):2067-85. Epub 2008 Jul 25. PMID: 18662169; DOI: 10.1111/j.1475-6773.2008.00882.x

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Table 2 shows the most recent reliability testing for PSI 04.

Table 2. Signal-to-Noise Ratio by Hospital Size Decile for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) in 2-year Pooled Data (2012-2013)

Hospital Size Decile	Number of Hospitals	Avg. Number of Discharges per Hospital in Decile	Avg. Signal-to-Noise Ratio for Hospitals in Decile
1 (smallest)	319	5.0	0.0579
2	320	11.9	0.1063
3	320	22.9	0.1732
4	320	37.4	0.2279
5	320	56.2	0.3094
6	320	81.6	0.3954
7	320	113.8	0.4738
8	320	157.2	0.5582
9	320	226.5	0.6464
10 (largest)	320	437.5	0.7765
Overall	3,199	115.0	0.6040

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2012 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.e., *what do the results mean and what are the norms for the test conducted?*)

For all-payer populations and across all hospitals in the AHRQ QI POA reference population, the overall signal to noise ratio for this indicator is moderate to good with an overall signal-to-noise ratio of 0.60. Hospitals with more than 82 qualifying discharges on average have risk adjusted rates with moderate to high reliability (average signal-to-noise ratio of 0.40 to 0.78). Signal-to-noise ratios were smaller for hospitals with fewer than approximately 82 qualifying discharges per year (average signal-to-noise ratio less than 0.40). Smoothed rates, which are recommended for all hospitals (and are implemented in the AHRQ software), address reliability concerns particularly for small hospitals.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

In this section, in addition to the empirical analyses completed, we summarize the most relevant literature, and provide a full evidence summary in the attached Evidence Form.

Empirical Validity Analyses

As part of the original testing and NQF endorsement process in 2007, we collaborated with Silber et al. to compare PSI 04 with other specifications of the failure-to-rescue concept, including related measure NQF 0352 ("Failure to Rescue In-Hospital Mortality (risk adjusted)").[3] This analysis was based on Medicare inpatient fee-for-service claims for general surgical admissions from July 1, 1999 through June 30, 2000, linked to the 2000 American Hospital Association Annual Survey. This data set included 1,467 hospitals and 403,679 Medicare beneficiaries between 65 and 90 years of age. To assess construct validity, we estimated logit models using detailed patient characteristics and 5 hospital characteristics shown to be associated with better quality of care in previous studies: (1) teaching status (member of the Council of Teaching Hospitals); (2) high technology status (open heart surgery or organ transplantation); (3) size greater than 200 beds; (4) bed-to-nurse ratio (RN plus LPN FTE positions); and (5) nursing skill mix ratio [RN/(RN + LPN)]. We report both "marginal" and "partial" results for each regression; marginal analyses used one hospital characteristic at a time along with all patient characteristics, whereas partial analyses adjusted for all hospital and patient variables simultaneously. Finally, the omega statistic represents the ratio of the squared sum of the log odds for patient characteristics at the discharge-level variables divided by the corresponding quantity for hospital-level variables. All else equal, outcome measures that have lower omega ratios may be more desirable quality indicators, because the lower the omega, the greater the hospital's impact on the outcome relative to the patient's impact.

In more recent analyses, we confirmed the association between teaching status and risk-adjusted PSI 04 rates using the 2012 and 2013 HCUP SID reference data set described above. These analyses used a broader definition of teaching status, as implemented in the HCUP program: "a hospital is considered to be a teaching hospital if it has an American Medical Association (AMA)-approved residency program, is a member of the Council of Teaching Hospitals, or has a ratio of full-time equivalent interns and residents to beds of .25 or higher."

Systematic Assessment of Face Validity

We utilized a structured panel review to evaluate face validity (from a clinical perspective) of the Patient Safety Indicators. The panels were convened in 2002. It is anticipated that the results of face validity review would be similar if panels were convened in more recent years, given that the clinical characteristics of these events, treatment and prevention approaches, and sequelae have not changed substantially since 2002. The clinical panel review process was based on the RAND appropriateness method, a modified Delphi process also known as a nominal group technique.

Twenty-one professional clinical organizations were invited to submit nominations. These organizations were selected based on the applicability of the specialty or subspecialty to potential Patient Safety Indicators. Clinical areas represented by the panels included internal medicine, cardiology, radiology, geriatrics, surgical and critical care nursing, anesthesiology, pharmacy, inpatient medicine and surgery (including thoracic, neurology, orthopedic, colorectal, urology, spine, and transplant surgical subspecialties). For assignments to each panel, a list of applicable specialties was identified for the indicators to be evaluated by that panel. Panelists were selected so that each panel had diverse membership in terms of practice characteristics and setting. For PSI 04, 7 members of a multispecialty panel completed the evaluation in full. Additional details of panel composition are available online at http://archive.ahrq.gov/clinic/tp/hospdatp.htm.

Panelists completed a 10-item questionnaire, tailored to each specific indicator. Following the initial rating of the indicators, panelists participated in a moderated 90-minute conference call, where opinions about the indicators were discussed. The panelists then completed the same 10-item questionnaire again, and submitted their final ratings. Ratings were summarized in accordance with the RAND Appropriateness Method.[4]

---

[3] Silber JH, Romano PS, Rosen AK, Wang Y, Ross RN, Even-Shoshan O, Volpp K. Failure-to-rescue: Comparing definitions to measure quality of care. Med Care 2007; 45:918-925.

[4] McDonald KM, Romano PS, Geppert J, Davies SM, Duncan BW, Shojania KG. Measures of Patient Safety Based on Hospital Administrative Data: The Patient Safety Indicators. Technical Review Number 5. Rockville, MD: Agency for Healthcare Research and Quality, 2002

For reference, the following text in black was previously submitted to NQF. Most of the information is outdated. We restricted our analysis to 20 states (4) for which HCUP State Inpatient Databases (SID) were available. There were 1,601 nonfederal, urban, general hospitals in those 20 states. Over 300 hospitals were eliminated from the sample because of key missing variables in the American Hospital Association (AHA) Annual Survey of Hospital data, which was also used for this study, or because they had missing observations for some of the OIs that we used. Thus, our sample consisted of 1,290 urban, acute-care hospitals for which complete data were available for 2001. [1]

The Agency for Healthcare Research and Quality Patient Safety Indicators (PSIs) were used to identify 14 PSIs among 161,004 surgeries. [5]

A likelihood ratio test of the hypothesis that the coefficients on all of these variables were equal to 0 (lambda) = 35.3, p< .01). [1]

We used propensity score matching and multivariate regression analyses to predict expenditures and outcomes attributable to the 14 PSIs. [5]

PSI 4 A higher risk-adjusted mortality rate for death among surgical inpatients with serious treatable complications is associated with significantly higher costs. The AHRQ QIs have the advantage of taking the multidimensional nature of hospital quality into account. As the coefficients on the AHRQ QIs show, measures of hospital quality can have conflicting effects on hospital costs. A single measure that combines these effects into one variable offers less insight into hospital performance than the outcomes for each measure.[1]

Principal Findings. Excess 90-day expenditures likely attributable to PSIs ranged from $646 for technical problems (accidental laceration, pneumothorax, etc.) to $28,218 for acute respiratory failure, with up to 20 percent of these costs incurred postdischarge. With a third of all 90-day deaths occurring postdischarge, the excess death rate associated with PSIs ranged from 0 to 7 percent. The excess 90-day readmission rate associated with PSIs ranged from 0 to 8 percent. Overall, 11 percent of all deaths, 2 percent of readmissions, and 2 percent of expenditures were likely due to these 14 PSIs. Conclusions. The effects of medical errors continue long after the patient leaves the hospital. Medical error studies that focus only on the inpatient stay can underestimate the impact of patient safety events by up to 20-30 percent. [5]

References

[1] Laditka JN, Laditka SB, Cornman CB. Evaluating hospital care for individuals with Alzheimer´s disease using inpatient quality indicators. Am J Alzheimers Dis Other Demen. 2005 Jan-Feb;20(1):27-36. PMID: 15751451.

[5] Encinosa WE, Hellinger FJ. The impact of medical errors on ninety-day costs and outcomes: an examination of surgical patients. Health Serv Res. 2008 Dec;43(6):2067-85. Epub 2008 Jul 25. PMID: 18662169; DOI: 10.1111/j.1475-6773.2008.00882.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Performance Measure Score

As summarized in the Evidence Form, numerous studies have linked failure to rescue measures, including PSI 04, to structure and process measures. Multiple studies have found lower FTR rates in hospitals with higher nurse-to-bed ratios[5,6,7,8,9,10,11,12], better nurse skill mix ratios[13,14,15,16,17,18,19], and better US-trained nurse ratios.[20]

[5] Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. JAMA 2002; 288:1987-1993.

[6] Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-staffing levels and the quality of care in hospitals. N Engl J Med 2002; 346:1715-1722.

[7] Friese CR, Aiken LH. Failure to rescue in the surgical oncology population: implications for nursing and quality improvement. Oncol Nurs Forum 2008; 35:779-785. PMC2562164.

[8] Ghaferi AA, Osborne NH, Birkmeyer JD, Dimick JB. Hospital characteristics associated with failure to rescue from complications after pancreatectomy. J Am Coll Surg 2010; 211:325-330.

[9] Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. Med Care 2013; 51:722-730.

Higher hospital volume was associated with lower FTR rates in at least 6 studies.[21] In addition, studies have found that hospitals with the highest patient satisfaction scores and hospitals with better compliance with NQF Safe Practices had lower risk adjusted odds of FTR.[22,23]

Empirical Validity Analyses

In the marginal analysis, we report the odds ratio and p value for each structural measure of quality when all patient characteristics are included in the model with only one hospital characteristic at a time. In the partial analysis, we report the same odds ratios and p values, using all patient discharge-level and hospital variables simultaneously. The latter approach may be more difficult to interpret due to collinearities among hospital characteristics. Teaching hospitals demonstrated odds ratios of 0.807 and 0.852 in marginal and partial analyses, respectively ($p<0.0001$ for both).

High technology hospitals demonstrated odds ratios of 0.924 ($p<0.005$) and 1.049 (NS) in marginal and partial analyses, respectively.

Large hospitals (>200 beds) demonstrated odds ratios of 0.872 ($p<0.0001$) and 0.917 ($p<0.01$) in marginal and partial analyses, respectively.

Less well staffed hospitals (with one additional bed per licensed nurse FTE) demonstrated odds ratios of 1.108 ($p<0.005$) and 1.044 (NS) in marginal and partial analyses, respectively.

Hospitals with better nursing skill mix (100% RN) demonstrated odds ratios of 0.832 and 0.870 in marginal and partial analyses, respectively ($p<0.0001$ for both).

The omega ratio summarizing the contribution of patient characteristics at the discharge-level versus hospital-level variables for PSI 04 was 57, compared with omega ratios of 189 for the overall risk-adjusted surgical mortality rate and 128 for NQF 0352.

We used all-payer data from 34 states in 2012-2013, described above, to confirm the association between hospital teaching status and lower PSI 04 rates. In these analyses, a much broader definition of teaching status was used, capturing not just COTH members but all hospitals with approved residency programs or more than

---

[10] Clarke SP, Aiken LH. Failure to rescue. Am J Nurs 2003; 103:42-47

[11] Schmid A, Hoffman L, Happ MB, Wolf GA, DeVita M. Failure to rescue: a literature review. J Nurs Adm 2007; 37:188-198.

[12] Park SH, Blegen MA, Spetz J, Chapman SA, De Groot H. Patient turnover and the relationship between nurse staffing and patient outcomes. *Res Nurs Health.* 2012;35(3):277-288.

[13] Blegen, M. A., et al. (2013). "Baccalaureate education in nursing and patient outcomes." J Nurs Adm 43(2): 89-94.

[14] Kendall-Gallagher D., Aiken L.H., Sloane D.M., and Cimiotti J.P.: Nurse specialty certification, inpatient mortality, and failure to rescue. J Nurs Scholarsh 2011; 43: pp. 188-194

[15] Aiken LH, Clarke SP, Cheung RB, Sloane DM, Silber JH. Educational levels of hospital nurses and surgical patient mortality. JAMA 2003; 290:1617-1623. PMC3077115.

[16] Silber JH, Romano PS, Rosen AK, Wang Y, Even-Shoshan O, Volpp KG. Failure-to-rescue: comparing definitions to measure quality of care. *Med Care.* 2007;45(10):918-925.

[17] Kendall-Gallagher D, Aiken LH, Sloane DM, Cimiotti JP. Nurse specialty certification, inpatient mortality, and failure to rescue. J Nurs Scholarsh 2011; 43:188-194. PMC3201820.

[18] Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. Med Care 2013; 51:722-730.

[19] Seago JA, Williamson A, Atwood C. Longitudinal analyses of nurse staffing and patient outcomes: More about failure to rescue. J Nurs Adm 2006; 36:13-21.

[20] Neff, D. F., et al. (2013). "Utilization of non-US educated nurses in US hospitals: implications for hospital mortality." Int J Qual Health Care 25(4): 366-372.

[21] Needleman J, Buerhaus P, Mattke S, Stewart M, Zelevinsky K. Nurse-staffing levels and the quality of care in hospitals. N Engl J Med 2002; 346:1715-1722.

[22] Sacks GD, Lawson EH, Dawes AJ, et al. Relationship Between Hospital Performance on a Patient Satisfaction Survey and Surgical Quality. *JAMA Surg.* 2015;150(9):858-864.

[23] Brooke BS, Dominici F, Pronovost PJ, Makary MA, Schneider E, Pawlik TM. Variations in surgical outcomes associated with hospital compliance with safety practices. *Surgery.* 2012;151(5):651-659.

0.25 residents per bed. In unadjusted analyses, teaching hospitals demonstrated a risk ratio of 1.131 (12.29% versus 10.86%). In adjusted analyses, using the current V6 risk-adjustment model, this risk ratio reversed to 0.976 (11.46% versus 11.74%). In adjusted analyses, using the proposed V7 risk-adjustment model, this risk ratio further improved to 0.975 (11.45% versus 11.75%).

Systematic Assessment of Face Validity

The multi-specialty Panel and Surgical Panel both rated the indicator as acceptable on overall usefulness as an indicator of quality of care.

Table 4. Clinician Panel Evaluations of the Face Validity for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) [4]

Multi-specialty Panel (MSP) Evaluation		
Overall Rating[1]	**Agreement[2]**	**Acceptability[3]**
7	Indeterminate	Acceptable

[1]Median panel overall rating of the indicator on a scale from 1 to 9, with the higher rating indicating better measurement
[2]Level of agreement, where "agreement" corresponds to little dispersion of opinion, "indeterminate" means that the opinion ranged but did not reach the point of clear "disagreement", the final category where there were panelists with diametrically different opinions
[3]"Acceptable" indicates that the indicator was rated as useful by almost all panelists. "Acceptable (-)" indicates that the indicator was rated as useful by most panelists, although a few rated it as less useful (but not as poor). "Unclear" indicates that panelists rated the usefulness of the indicator as moderate. For further details of methods, see http://archive.ahrq.gov/clinic/tp/hospdatp.htm
[4]PSI 04 was evaluated under a previous name (i.e. Failure to Rescue).

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The consistent associations with structural measures of hospital quality, including higher nurse staffing, better nursing skill mix, higher hospital volume (beds), and teaching status, suggest that PSI 04 is a valid measure. These findings are supported by other studies (summarized in the Evidence Form) that showed lower PSI 04 or failure-to-rescue rates at hospitals with better patient satisfaction and higher adherence to NQF Safe Practices. Teaching hospitals had higher unadjusted PSI 04 rates, but lower adjusted PSI 04 rates, relative to nonteaching hospitals. However, this effect was less pronounced with a more inclusive definition of teaching hospitals (and all-payer data instead of Medicare data).

PSI 04 has acceptable face/content validity based on clinical panel evaluation.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

Empirical Evaluation of Exclusions: Using the 2013 data from 34 states, we examined the percent of potential denominator cases excluded by each criterion as listed in the measure specifications.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

Table 5 shows the results of the most recent exclusions analysis.

Table 5. Number and Percent of Discharges Excluded, by Denominator Exclusion Criteria, for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) [1]

PSI 04	Denominator			Numerator		
Exclusion Name	**Exclusion Count**	**After Exclusions**	**% Change**	**Exclusion Count**	**After Exclusions**	**% Change**
No Exclusions applied		**329,716**			**37,698**	
Exclude Transfers to an acute care facility	9,889	319,827	3.0%	0	37,698	0.0%
Stratum: PE/DVT						
Exclude principal dx PE/DVT	884	328,832	0.3%	7	37,691	0.0%
Exclude principal dx Abortion-Related and Postpartum Obstetric Pulmonary Embolism	6	329,710	0.0%	0	37,698	0.0%
Stratum: Pneumonia						
Exclude principal dx Pneumonia	184	329,532	0.1%	12	37,686	0.0%
Exclude principal dx respiratory Complications	126	329,590	0.0%	0	37,698	0.0%
Exclude Diagnosis Viral Pneumonia	343	329,373	0.1%	17	37,681	0.0%
Exclude Dx of Immunocompromised	8,696	321,020	2.6%	632	37,066	1.7%
Exclude Procedure of Immunocompromised	172	329,544	0.1%	1	37,697	0.0%
Exclude MDC 4	9,006	320,710	2.7%	428	37,270	1.1%
Exclude Lung Cancer Procedure	47	329,669	0.0%	4	37,694	0.0%
Stratum Sepsis						
Exclude principal dx Septicemia	45,202	284,514	13.7%	3,381	34,317	9.0%
Exclude principal dx of Infection	14,982	314,734	4.5%	843	36,855	2.2%
Exclude Diagnosis of Immunocompromised	4,830	324,886	1.5%	1,050	36,648	2.8%
Exclude Procedure of Immunocompromised	44	329,672	0.0%	5	37,693	0.0%
Exclude Length of Stay	1,476	328,240	0.4%	703	36,995	1.9%

Less than 4						
Stratum: Shock/Cardiac Arrest						
Exclude principal dx Shock	99	329,617	0.0%	55	37,643	0.1%
Exclude principal dx Trauma	0	329,716	0.0%	0	37,698	0.0%
Exclude principal dx Hemorrhage	453	329,263	0.1%	81	37,617	0.2%
Exclude principal dx GI Hemorrhage	3,620	326,096	1.1%	2,239	35,459	5.9%
Exclude principal dx Abortion-Related Shock	947	328,769	0.3%	175	37,523	0.5%
Exclude MDC 4 or 5	27,819	301,897	8.4%	6,611	31,087	17.5%
Stratum GI hemorrhage / Acute ulcer						
Exclude principal dx GI Hemorrhage-Acute Ulcer	714	329,002	0.2%	12	37,686	0.0%
Exclude principal dx Blood Loss / Anemia	141	329,575	0.0%	0	37,698	0.0%
Exclude principal dx Trauma	3,117	326,599	0.9%	78	37,620	0.2%
Exclude principal dx Alcoholism	354	329,362	0.1%	11	37,687	0.0%
Exclude MDC 6 or 7	14,134	315,582	4.3%	128	37,570	0.3%
All Exclusions applied		**182,429**			**21,226**	

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

[1]This indicator does not have numerator exclusion criteria.

[2]Potential numerator cases are those that would have qualified for the numerator if not for a particular denominator exclusion criterion.

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

The stratum-specific exclusions are meant to exclude cases for which the "complication" was actually the principal reason for admission or the primary indication for surgery. Some exclusions (e.g. immunocompromised state) are intended to exclude patients for whom death may be the expected outcome (i.e., less preventable). For example, patients presenting with acute hemorrhagic shock due to ongoing blood loss are unlikely to survive.

All patients transferred to other hospitals must be excluded from the analysis because the relevant outcome of these patients (i.e., dead or alive at the time of discharge from the acute inpatient setting) cannot be ascertained without social security numbers or other data elements to support linkage.

Although many of these exclusions are rare, they ensure face validity. Note that AHRQ does NOT exclude patients with complications that were present on admission (i.e., upon transfer from another hospital, emergency department, or ambulatory surgery center). This decision was based on Needleman et al.'s analysis, showing

that PSI 04 mortality was higher among patients with hospital-acquired complications than among patients with present-on-admission complications (22% versus 13%), and that nurse staffing (licensed nurse hours per bed) and skill mix were highly associated with PSI 04 rates, regardless whether complications present on admission were included or excluded.[24]

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 362 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Clinical Factors

For each PSI 04 stratum, we considered a standard set of covariates grouped into four categories: demographics, severity of illness, comorbidities and transfer-in status. Covariates that were considered as potential risk adjusters included gender and age (in mutually exclusive 5-year age categories), Major Diagnostic Categories (MDCs), Modified Diagnostic Related Groups (MDRGs) (defined as the base MS-DRG without comorbidity or complication distinctions), AHRQ Comorbidity Software categories, and whether the patient was transferred in from another facility. Only those covariates present in at least 30 records for that PSI 04 strata are retained. A parsimonious model was identified using backward stepwise selection with bootstrapping.

The omitted covariate within mutually exclusive categories is the reference group for those categories. Reference categories are usually 1) the most common and/or 2) the least risk. The choice of omitted reference category does not affect predicted probabilities or model performance.

For the MDRGs, the risk reported is the residual risk after adjustment for the MDC to which the MDRG belongs. Likewise, the risk reported for MDCs represents the average risk of all MSDRGs in that MDC not included in the model.

The risk-adjusted rate for the overall PSI 04 is calculated as the observed to expected ratio multiplied by the reference population rate, where the observed and expected values are summed across five strata (categories) of PSI 04 risk. This approach differs from other AHRQ Patient Safety Indicators without strata, in that each discharge-record's expected value is computed using one of five distinct stratum-specific risk adjustment models that correspond to an assigned PSI 04 stratum. The five PSI 04 strata group records together based on secondary diagnoses that represent complications of care, and place the patient at risk of death (which is the numerator of PSI 04).

Additional details are available in the *AHRQ Quality Indicator Empirical Methods* document, included in the supplemental file and available on the AHRQ QI website.

[24] Needleman J, Buerhaus PI, Vanderboom C, Harris M. Using present-on-admission coding to improve exclusion rules for quality metrics: the case of failure-to-rescue. *Med Care.* 2013;51(8):722-730.

Sociodemographic Factors

There is no evidence or causal model to suggest that socioeconomic factors are associated with death following serious surgical complications independent of quality of care, or are mediated by pre-hospital care (which may not fall within the proper realm of hospital accountability). Accordingly, consistent with the guidance provided by NQF in the SDS Trial Period FAQs, AHRQ believes that it would be inappropriate to include other SDS variables in the risk-adjustment approach for PSI 04, which is an in-hospital outcome measure.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

This section includes a summary of the selected risk factors for each stratum, used together to construct the risk model for the overall PSI 04 measure. Details of the current risk adjustment coefficients for each PSI 04 stratum can be found in the attached technical specifications.

STRATUM_SHOCK: The risk model includes 75 risk categories, including 24 age-gender categories in 5-year age categories between ages 30 and 89, and 2 age-gender categories below age 30 (i.e. 18-29), transfer in from another acute care facility and 14 comorbidities. The remainder of selected risk factors account for the reason for admission and the type of surgery that was performed during the hospitalization, including MDC and MS-DRGs collapsed to remove Complication or Comorbidity/ Major Complication or Comorbidity (CC/MCC) distinctions.

STRATUM_SEPSIS: The risk model includes 81 risk categories, including 24 age-gender categories in 5-year age categories between ages 30 and 89, and 2 age-gender categories below age 30 (i.e. 18-29), transfer in from another acute care facility and 18 comorbidities. The remainder of selected risk factors account for the reason for admission and the type of surgery that was performed during the hospitalization, including MDC and MS-DRGs collapsed to remove Complication or Comorbidity/ Major Complication or Comorbidity (CC/MCC) distinctions.

STRATUM_PNEUMONIA: The risk model includes 89 risk categories, including 24 age-gender categories in 5-year age categories between ages 30 and 89, and 2 age-gender categories below age 30 (i.e. 18-29), transfer in from another acute care facility and 22 comorbidities. The remainder of selected risk factors account for the reason for admission and the type of surgery that was performed during the hospitalization, including MDC and MS-DRGs collapsed to remove Complication or Comorbidity/ Major Complication or Comorbidity (CC/MCC) distinctions.

STRATUM_DVT: The risk model includes 56 risk categories, including 24 age-gender categories in 5-year age categories between ages 30 and 89, and 2 age-gender categories below age 30 (i.e. 18-29), transfer in from another acute care facility and 10 comorbidities. The remainder of selected risk factors account for the reason for admission and the type of surgery that was performed during the hospitalization, including MDC and MS-DRGs collapsed to remove Complication or Comorbidity/ Major Complication or Comorbidity (CC/MCC) distinctions.

STRATUM_GI_HEM: The risk model includes 61 risk categories, including 24 age-gender categories in 5-year age categories between ages 30 and 89, and 2 age-gender categories below age 30 (i.e. 18-29), transfer in from another acute care facility and 15 comorbidities. The remainder of selected risk factors account for the reason for admission and the type of surgery that was performed during the hospitalization, including MDC and MS-DRGs collapsed to remove Complication or Comorbidity/ Major Complication or Comorbidity (CC/MCC) distinctions.

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Not applicable (see above)

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

This analysis evaluates the performance of the risk adjustment model(s) with respect to the event of interest (i.e., in-hospital death). The measure of discrimination, how well the risk adjustment model distinguishes events from non-events, is the c-statistic (also known as the area under a receiver operating characteristic curve). The c-statistic is computed by assigning each observation a predicted probability of the outcome from the risk-adjustment model, based on the value of the observed covariates and the parameter estimates from the risk-adjustment model. Two copies of the dataset are sorted, first from highest to lowest predicted probability and second from lowest to highest predicted probability. Random sampling is used to create a set of paired observations. Pairs that consist of one event and one non-event (discordant pairs) are kept and concordant pairs are discarded. The c-statistic represents the proportion of discordant pairs of observations for which the observation with the event had a higher predicted probability from the risk-adjustment model than the observation without the event. C-statistics above 0.70 and below 0.80 have moderate discrimination. Above 0.80, the discrimination is considered high. We did not employ common "goodness of fit" tests because these tests tend to be uninformative with large samples.

We also evaluated the calibration of the risk adjustment model by evaluating how closely observed and predicted rates compare across deciles of the predicted rate. This analysis splits the sample into deciles based on predicted rates, and then compares these rates with the observed rates for the population in each decile. A well calibrated model, or one that does not over or under-estimate risk, will have comparable observed and predicted rates across the risk spectrum.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to ***2b4.9***

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

The c-statistic for the overall PSI 04 model is 0.829 in the 2013 HCUP data (described above).

Note that there are actually five distinct risk models; one for each type of complication or stratum. These five models currently have c statistics that range from 0.726 to 0.860. Enhancements now being tested for version 7 (i.e., adjusting for the severity and timing of the triggering complication, as well as all of the factors listed in 2b4.4a above) will increase these c statistics to 0.779 to 0.877. The overall c statistic reported in above represents the overall performance of all five models.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

Table 6. Risk adjustment Model Discrimination and Calibration for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04), 2013

Predicted Rate Decile	Number of Discharges per Decile	Predicted Rate (per 1,000 surgical discharges)	Observed Rate (per 1,000 surgical discharges)	Observed to Predicted Ratio
1 (lowest)	18,251	5.4184	3.0683	0.57
2	18,251	13.9581	11.4514	0.82
3	18,251	23.4806	17.4785	0.74
4	18,251	34.4656	34.3543	1.00
5	18,251	49.3254	53.2026	1.08
6	18,252	70.4652	71.8825	1.02

7	18,251	102.9786	108.9255	1.06
8	18,251	157.9172	162.8952	1.03
9	18,252	248.7309	253.7804	1.02
10 (highest)	18,251	457.1238	446.8248	0.98

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
See calibration by decile in Table 6 in 2b4.7

2b4.9. Results of Risk Stratification Analysis:

Not applicable

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

The risk-adjustment model has moderately high discrimination, based on a c statistic of 0.829 (i.e., in 83% of randomly selected pairs of discordant observations, the patient who experienced PSI 04 had a higher probability of experiencing the event than the patient who did not). A model that is well calibrated will have observed values similar to predicted values across the predicted value deciles. This indicator is well calibrated, as the observed to predicted ratio values across the deciles range between 0.74 to 1.08 for all deciles except the lowest decile. For patients with very low predicted rates, the relative difference between observed and predicted values is greater, but this is not particularly concerning due to the very small number of events that occur in this risk stratum.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

This analysis assesses the probability that a hospital is higher or lower than a benchmark or threshold, given hospital size. It reflects whether the indicator can discriminate the best performing hospitals from the lower performing hospitals.

For this analysis, "benchmark" refers to the smoothed indicator rate based on the 20th percentile of the reference population (i.e., 20% of hospitals have a lower mortality rate or better performance). "Threshold" refers to the indicator rate based on the 80th percentile (i.e., 80% have lower mortality or better performance). Assuming an underlying Gamma distribution for the smoothed rates of the measure, the benchmark and threshold values are identified using population reference rates and signal variances computed from the entire AHRQ QI POA Reference Population. Hospital-level 90% confidence limits for smoothed rates are also computed from the Gamma distribution.

The analysis is reported by size decile, based on the denominator cases, demonstrating performance across hospitals of various sizes. Each hospital is assumed to have an underlying distribution of smoothed rates that follows a Gamma distribution. The parameters of a Gamma distribution are shape and scale. For each hospital the shape is calculated as ((smoothed rate)^2/ smoothed rate variance), and the scale is calculated as (smoothed rate variance / smoothed rate). The smoothed rate variance (aka posterior variance) is calculated as the signal variance – (reliability weight * signal variance). The reliability weight is calculated as (signal variance / (signal variance + noise variance)). Hospitals are

ranked by size and grouped into 10 equal categories of size (deciles). The Benchmark and Threshold are compared to the Gamma distribution of the smoothed rates for each hospital to determine if the hospital rate is better or worse than the Benchmark and Threshold rates with 95% probability. This provides a 95% confidence interval for the Benchmark and Threshold rate.

Table 7 reports the proportion of hospitals above (better than) and below (worse than) the Benchmark and Threshold rates and the proportion not classified as either above or below. The hospitals not classified as either better or worse have rates that fall within the 95% confidence interval.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Table 7. Performance Categories Using Smoothed Hospital Rates by Hospital Size Decile for Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI 04) in 2-year Pooled Data (2012-2013)

			Benchmark			Threshold		
Hospital Size Decile	Number of Hospitals	Average Number of Denominator Discharges Per Hospital	Proportion Better	Proportion Worse	Proportion Unclassified	Proportion Better	Proportion Worse	Proportion Unclassified
(smallest) 1	319	5.0	0.0000	0.0063	0.9937	0.0000	0.0000	1.0000
2	320	11.9	0.0000	0.0219	0.9781	0.0000	0.0000	1.0000
3	320	22.9	0.0000	0.0938	0.9063	0.0125	0.0000	0.9875
4	320	37.4	0.0000	0.1281	0.8719	0.0156	0.0000	0.9844
5	320	56.2	0.0000	0.2188	0.7813	0.0625	0.0031	0.9344
6	320	81.6	0.0000	0.3281	0.6719	0.1344	0.0000	0.8656
7	320	113.8	0.0000	0.3375	0.6625	0.2344	0.0125	0.7531
8	320	157.2	0.0031	0.4094	0.5875	0.2344	0.0031	0.7625
9	320	226.5	0.0031	0.4500	0.5469	0.3781	0.0188	0.6031
10 (largest)	320	437.5	0.0094	0.6188	0.3719	0.4344	0.0438	0.5219
Overall	3,199	115.0	0.0016	0.2613	0.7371	0.1507	0.0081	0.8412

Source: HCUP State Inpatient Databases (SID). Healthcare Cost and Utilization Project (HCUP). 2012 - 2013. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov (AHRQ QI Software Version 6.0)

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

As hospital size increases, the discrimination also increases such that for hospitals in the largest 2 deciles the algorithm classifies 40% - 48% of hospitals against the threshold and 45%-63% of hospitals against the benchmark, based on conventional statistical criteria. Over all hospitals, using smoothed rates, this indicator has limited discrimination for identifying low or high performing hospitals; 16% of hospitals can be classified as better or worse than the threshold (the percentage classified as either above or below the threshold) and 27% as better or worse than the benchmark (the percentage classified as either above or below the benchmark), based on conventional statistical criteria. In this example, use of smoothed rates "shrinks" the performance distribution across hospitals, which typically decreases performance discrimination. Although this means that hospitals with few inpatient post-surgical complications (the PSI 04 denominator) cannot be identified as low

or high performers unless their PSI 04 rates vary widely from the benchmark/threshold, this shrinkage approach also makes the measure less likely to classify hospitals as low performing when they are not.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS 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.***

Not applicable

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
Not applicable

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
Not applicable

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
Not applicable

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

The AHRQ QIs use frequently reported administrative data variables. PSI 04 excludes cases with missing discharge disposition, age, sex, discharge quarter, discharge year, and principal diagnosis. These variables are required for indicator construction and are required of all hospital discharge records. The frequency of missing data for each variable is available by state and year from the AHRQ HCUP website (http://www.hcup-us.ahrq.gov/cdstats/cdstats_search.jsp).

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

For these variables, frequencies of missing data are typically less than 1% of the state database. It is unlikely that bias would occur from such a low frequency of missing data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Exclusion of cases with missing data for these variables is appropriate.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
Because the indicator is based on readily available administrative billing and claims data and U.S. Census data, feasibility is not an issue.
The AHRQ QI software has been publicly available at no cost since 2001; Users have over ten years of experience using the AHRQ QI software in SAS and Windows.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees. Software is freely available from the AHRQ Quality Indicators website (http://www.qualityindicators.ahrq.gov/).

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Arizona Department of Health Services, AZ Hospital Compare, MONAHRQ website http://pub.azdhs.gov/hospital-discharge-stats/2012/AboutQualityRatings.html CareChex (Division of Quantros) http://www.carechex.com/QualityIndicators.aspx CMS Medicare Hospital Compare Program https://www.medicare.gov/hospitalcompare/Data/Measures-Displayed.html# Commonwealth Fund, Why Not the Best http://whynotthebest.org/methodology Connecticut Department of Health Services, CT Hospital Compare, MONAHRQ website http://ctmonahrq.ct.gov/2012/index.html#/resources/AboutQualityRatings Connecticut Hospital Association http://www.cthosp.org/advocacy/quality-and-patient-safety/hospital-quality-reporting-website/ Consumer Reports http://www.consumerreports.org/health/resources/pdf/how-we-rate-hospitals/How%20We%20Rate%20Hospitals.pdf HealthGrades https://d2dcgio3q2u5fb.cloudfront.net/54/98/f79cdfd84640a03792ea092f20a8/2014-patient-safety-methodology.pdf Hospital Safety Score http://www.hospitalsafetyscore.org/media/file/HospitalSafetyScore_ScoringMethodology_Spring2015_Final.pdf Iowa Healthcare Collaborative https://iowareport.ihconline.org/Public/Reports.aspx?FID=778&F1ID=0&F2ID=0&F3ID=0&CID=2&PID=4 Kentucky Cabinet for Health and Family services https://prd.chfs.ky.gov/MONAHRQ/2012/MONAHRQ/AboutQualityRatings.html Kentucky Hospital Association Quality Data http://info.kyha.com/qualitydata/psisite/SelectPSIReport.asp?IndID=PSI4&TimePeriod=5&GroupOpt=none&SortOrder=hospital&SortDir=ASC Louisiana Hospital Inform http://lahospitalinform.org/index.html Maryland Health Care Commission, MONAHRQ Website http://www.hscrc.state.md.us/documents/md-maphs/wg-meet/di/2014-03-04/MHCC%20Inpatient%20Measures%20Inventory%20QBR%20highlights.pdf Maine Health Data Organization (MHDO), MONAHRQ Website https://mhdo.maine.gov/monahrq/#/resources/AboutQualityRatings Minnesota Community Measurement http://mncm.org/wp-content/uploads/2014/02/2013-HCQR-Final-2.4.2014.pdf

	Nevada Compare Care, MONAHRQ website http://nevadacomparecare.net/MQ2014/index.html#/professional/resources/AboutQualityRatings Niagara Health Quality Coalition, New York State Hospital Report Card http://www.myhealthfinder.com/newyork15/main_byproc.php Norton Healthcare http://www.nortonhealthcare.com/QualityReport Oklahoma State Department of Health, MONAHRQ https://www.phin.state.ok.us/ahrq/MONAHRQ%202010/Methodology.html South Dakota Association of Healthcare Organizations http://www.sdhospitalquality.org/search.php http://healthdata.dshs.texas.gov/Hospital/PatientSafetyQualityIndicators Texas Department of State Health Services Texas Health Resources https://www.texashealth.org/Documents/System/Quality_Patient_Safety/Reports/03-02-2016_Surgery.pdf U.S. News and World Report http://www.usnews.com/pubfiles/BH2015-16MethodologyReport.pdf Utah Department of Health, MONAHRQ website https://health.utah.gov/myhealthcare/monahrq/ Virginia Health Information http://www.vhi.org/MONAHRQ/default.asp?yr=2013 Washington State, MONAHRQ website http://www.wamonahrq.net/MONAHRQ_5p0_WA_2012/index.html#/resources/AboutQualityRatings WHA Information Center (Wisconsin Hospital Association) http://www.whainfocenter.com/uploads/PDFs/Publications/QualityIndicators/2012_WI_IQIReport.pdf CMS Hospital Quality Initiative: Outcome Measures http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) CMS Hospital Compare http://www.medicare.gov/hospitalcompare/Data/Measures-Displayed.html University HealthSystem Consortium/Vizient https://www.vizientinc.com/clinical-analytics-and-benchmarking.htm Quality Improvement (Internal to the specific organization) BayCare https://baycare.org/quality-report-card/surgical-complications-0715 Blue Cross Blue Shield of North Carolina http://www.bcbsnc.com/content/providers/hqp/index.htm Greenville Health System, Quality and Safety Report http://www.ghs.org/upload/docs/Reports/2013-April-Quality-Report.pdf Northwestern Memorial Hospital, Patient Safety Indicator Monitoring Plan https://www.nm.org/location/northwestern-memorial-hospital/quality-nmh/view-our-quality-ratings-nmh/surgery-nmh/general-surgery-nmh Upstate University Hospital http://qoc.upstate.edu/QualityOfCare.cfm?quality_measure_group_id=7

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Public Reporting:

Arizona Department of Health Services, AZ Hospital Compare, MONAHRQ website
Hospital quality ratings from all hospitals in Arizona
http://pub.azdhs.gov/hospital-discharge-stats/2012/AboutQualityRatings.html

CareChex (Division of Quantros)
Provides comprehensive reports of hospitals to consumers, providers and purchasers
http://www.carechex.com/QualityIndicators.aspx

CMS Medicare Hospital Compare Program
Publically available database containing information about the quality of care at over 4,000 Medicare-certified hospitals across the U.S.
https://www.medicare.gov/hospitalcompare/Data/Measures-Displayed.html#

CMS Hospital Quality Initiative: Outcome Measures
Produces a chartbook of hospital outcome measures
http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html

Commonwealth Fund, Why Not the Best
Provides performance and quality ratings for most US hospitals
http://whynotthebest.org/methodology

Connecticut Department of Health Services, CT Hospital Compare, MONAHRQ website
Hospital quality ratings from all hospitals in Connecticut
http://ctmonahrq.ct.gov/2012/index.html#/resources/AboutQualityRatings

Connecticut Hospital Association
Provide quality of care for hospitals in Connecticut
http://www.cthosp.org/advocacy/quality-and-patient-safety/hospital-quality-reporting-website/

Consumer Reports
Hospital measure performance compared to external hospitals
http://www.consumerreports.org/health/resources/pdf/how-we-rate-hospitals/How%20We%20Rate%20Hospitals.pdf

HealthGrades
Healthgrades measures 40 million patient records from 4,500 hospitals nationwide for the most recent three-year period. Consumer-targeted hospital and provider ratings
https://d2dcgio3q2u5fb.cloudfront.net/54/98/f79cdfd84640a03792ea092f20a8/2014-patient-safety-methodology.pdf

Hospital Safety Score
PSI 04 is one component of a single composite score that represents a hospital's overall performance in patient safety
http://www.hospitalsafetyscore.org/media/file/HospitalSafetyScore_ScoringMethodology_Spring2015_Final.pdf

Iowa Healthcare Collaborative
Hospital quality ratings from hospitals in Iowa
https://iowareport.ihconline.org/Public/Reports.aspx?FID=778&F1ID=0&F2ID=0&F3ID=0&CID=2&PID=4

Kentucky Cabinet for Health and Family services
Hospital quality ratings from hospitals in Kentucky
https://prd.chfs.ky.gov/MONAHRQ/2012/MONAHRQ/AboutQualityRatings.html

Kentucky Hospital Association Quality Data
Hospital quality ratings from most hospitals in Kentucky
http://info.kyha.com/qualitydata/psisite/SelectPSIReport.asp?IndID=PSI4&TimePeriod=5&GroupOpt=none&SortOrder=hospital&SortDir=ASC

Louisiana Hospital Inform

Hospital quality ratings from hospitals in Louisiana
http://lahospitalinform.org/index.html

Maryland Health Care Commission, MONAHRQ Website
Collects and provides quality ratings on hospitals across Maryland
http://www.hscrc.state.md.us/documents/md-maphs/wg-meet/di/2014-03-04/MHCC%20Inpatient%20Measures%20Inventory%20QBR%20highlights.pdf

Maine Health Data Organization (MHDO), MONAHRQ Website
Hospital quality ratings from all hospitals in Maine
https://mhdo.maine.gov/monahrq/#/resources/AboutQualityRatings

Minnesota Community Measurement
Minnesota Community Measurement is a nonprofit healthcare data reporting organization. Provides quality ratings on hospitals across Minnesota.
http://mncm.org/wp-content/uploads/2014/02/2013-HCQR-Final-2.4.2014.pdf

Nevada Compare Care, MONAHRQ website
Hospital quality ratings from most hospitals in Nevada
http://nevadacomparecare.net/MQ2014/index.html#/professional/resources/AboutQualityRatings

Niagara Health Quality Coalition, New York State Hospital Report Card
Consumer focused public report of quality indicator performance for NY hospitals.
http://www.myhealthfinder.com/newyork15/main_byproc.php

Norton Healthcare
Report patient satisfaction scores in Norton Healthcare hospitals and their performance on nationally recognized quality indicators and practices http://www.nortonhealthcare.com/QualityReport

Oklahoma State Department of Health, MONAHRQ
Compares quality ratings on hospitals across Oklahoma
https://www.phin.state.ok.us/ahrq/MONAHRQ%202010/Methodology.html

South Dakota Association of Healthcare Organizations
Use PSI 04 in a composite of serious complications in report of South Dakota hospital quality.
http://www.sdhospitalquality.org/search.php

Texas Department of State Health Services
Texas Health Care Information Collection
http://healthdata.dshs.texas.gov/Hospital/PatientSafetyQualityIndicators

Texas Health Resources
Provides quality and safety reports for all Texas Health Resources
https://www.texashealth.org/Documents/System/Quality_Patient_Safety/Reports/03-02-2016_Surgery.pdf

U.S. News and World Report
National publication that lists ratings of U.S. medical centers based on performance
http://www.usnews.com/pubfiles/BH2015-16MethodologyReport.pdf

Utah Department of Health, MONAHRQ website
Report hospital quality for all hospitals in Utah
https://health.utah.gov/myhealthcare/monahrq/

Virginia Health Information
Compares quality ratings on hospitals across Virginia
http://www.vhi.org/MONAHRQ/default.asp?yr=2013

Washington State, MONAHRQ website
Information system of inpatient care utilization, quality, and potentially avoidable stays in Washington State's community hospitals
http://www.wamonahrq.net/MONAHRQ_5p0_WA_2012/index.html#/resources/AboutQualityRatings

WHA Information Center (Wisconsin Hospital Association)
Wisconsin Inpatient Hospital Quality Indicators Report
http://www.whainfocenter.com/uploads/PDFs/Publications/QualityIndicators/2012_WI_IQIReport.pdf

Quality Improvement (external benchmarking to multiple organizations):

CMS Hospital Compare
Publically available performance measures for hospitals
http://www.medicare.gov/hospitalcompare/Data/Measures-Displayed.html

University HealthSystem Consortium/Vizient
Internal quality improvement efforts, documentation, and evaluation of AHRQ PSIs for quality improvement by its members
https://www.vizientinc.com/clinical-analytics-and-benchmarking.htm

Quality Improvement (internal to the specific organization):

BayCare
Provide information on quality of hospital care within the BayCare health system
https://baycare.org/quality-report-card/surgical-complications-0715

Blue Cross Blue Shield of North Carolina
Stimulate improvements in quality and safety within hospitals
http://www.bcbsnc.com/content/providers/hqp/index.htm

Greenville Health System, Quality and Safety Report
All data was collected from four hospitals in the Greenville Health system and compared with internal rates
http://www.ghs.org/upload/docs/Reports/2013-April-Quality-Report.pdf

Northwestern Memorial Hospital, Patient Safety Indicator Monitoring Plan
Quality improvement initiative at 894-bed academic hospital
https://www.nm.org/location/northwestern-memorial-hospital/quality-nmh/view-our-quality-ratings-nmh/surgery-nmh/general-surgery-nmh

Upstate University Hospital
Report of hospital rates against national benchmark (published online)
http://qoc.upstate.edu/QualityOfCare.cfm?quality_measure_group_id=7

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
n/a

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
n/a

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in

use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

See Table 1 in response to question 1b.2 (also included in supplemental materials)

We observe that PSI 04 rates have been relatively stable from 2011-2013 in the AHRQ QI POA Reference Population data (116-118 deaths per 1000 patients with perioperative or postoperative complications). An earlier study of administrative data showed a decrease by 6.05% per year (p<0.0001) (Downey et al., 2012).

Downey, J. R., et al. (2012). "Is patient safety improving? National trends in patient safety indicators: 1998-2007." Health Serv Res 47(1 Pt 2): 414-430.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
No evidence has been identified suggesting unintended consequences for this measure.
Coding professionals follow detail guidelines, are subject to training and credentialing requirements, peer review and audit.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0352 : Failure to Rescue In-Hospital Mortality (risk adjusted)
0353 : Failure to Rescue 30-Day Mortality (risk adjusted)

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
n/a

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?
No

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
NQF 0353 uses 30-day mortality (dated from the date of the surgical admission), regardless of location, for the numerator. This is a different outcome from in-hospital mortality, and is only available in a very limited number of data sets, so NQF 0353 is a related (not competing) measure. NQF 0352 is a measure of in-hospital mortality, similar to PSI 04 (NQF 0351), but it has a different target population, so NQF 0352 is a related (not competing) measure. Specifically, the denominator for NQF 0352 and NQF 0353 is limited to surgical MS-DRGs in MDC 6 (Digestive System), MDC 7 (Hepatobiliary), MDC 9 (Skin, subcutaneous tissue, breast), MDC 10 (Endocrine, nutritional, metabolic), MDC 8 (Musculoskeletal and connective tissue), and MDC 5 (Circulatory system). By contrast, the denominator for PSI 04 (NQF 0351) also includes patients undergoing transplantation, neurosurgical, ophthalmologic, otolaryngologic (ENT), pulmonary/respiratory, urologic, gynecologic, hematologic, infection-related, trauma-related, and burn-related major procedures (if they otherwise qualify for the denominator). Therefore, the clinical/specialty breadth of the current measure is substantially greater than that of NQF 0352. Although all three of these measures are focused on "surgical patients between ages 18 and 90 admitted to an acute care hospital," the available risk-adjustment for NQF 0352 and NQF 0353 is based on Medicare fee-for-service claims data, which greatly limits the usefulness of these two measures for users with all-payer data sets (i.e., hospitals and hospital systems/associations, state and regional health data agencies, regional quality collaboratives and other "report card" sponsors, and researchers using HCUP or similar data). By contrast, the publicly available risk-adjustment for PSI 04 (NQF 0351) is based on all-payer data from 34 US states. The target population for PSI 04 (NQF 0351) is substantially broader than the target population for NQF 0352 and NQF 0353, as described above. Another key difference in denominator specifications is that PSI 04 (NQF 0351) only includes patients who experienced one or more of five broad categories of perioperative or postoperative complications, as defined by the strata. By contrast, the denominators of NQF 0352 and NQF 0353 include patients with a much wider set of 38 perioperative or postoperative complications. More importantly, in-hospital death after surgery automatically qualifies a patient for the denominator of NQF 0352, regardless whether the patient had any reported complication. As a result, the numerator of NQF 0352 includes ALL in-hospital deaths after eligible operations, whereas the numerator of PSI 04 (NQF 0351) only includes in-hospital deaths that follow one or more of the stratum-defining complications. Previous studies suggest that PSI 04 (NQF 0351) captures about 42-49% of all in-hospital deaths after qualifying operations, whereas NQF 0352 captures 100% of these deaths. The clinical rationale for this difference is that focusing on a narrower subset of deaths provides an easier target for quality improvement efforts and makes the indicator more sensitive to nursing-related quality of care (i.e., nurses are presumably less likely to be able to "rescue" patients from sudden unexpected deaths or "planned" deaths, in which physicians' orders and/or advance directives do not allow cardiopulmonary resuscitation or similar efforts). Specifically, a 2007 analysis cited in the Testing Form showed that the omega ratio summarizing the contribution of patient characteristics at the discharge-level versus hospital-level variables for explaining PSI04 (NQF 0351) was 57, compared with omega ratios of 189 for the overall risk-adjusted surgical mortality rate and 128 for NQF 0352. In other words, NQF 0352 is more heavily influenced by patient characteristics, whereas PSI 04 (NQF 0351) better isolates the hospital quality effect (albeit at the price of lower reliability, given that it only captures 42-49% of all in-hospital deaths after qualifying operations).

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** PSI04_Supplemental_file_160531.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Agency for Healthcare Research and Quality
Co.2 **Point of Contact:** Pamela, Owens, Pam.Owens@ahrq.hhs.gov, 301-427-1412-
Co.3 **Measure Developer if different from Measure Steward:** Agency for Healthcare Research and Quality
Co.4 **Point of Contact:** Mamatha, Pancholi, Mamatha.Pancholi@ahrq.hhs.gov, 301-427-1470-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
In 2002, a workgroup convened and provided feedback on key indicator development decisions and methodology, including the usefulness of the Death Rate among Surgical Inpatients with Serious Treatable Complications (PSI04), formerly known as Failure to Rescue (PSI04). The active members of the panel were:

Michael Barrett, MD, Internist and Cardiologist
Blue Bell, PA
Medical College of Pennsylvania Hospital
Nominated by the American College of Physicians

William Golden, MD, Professor of medicine, Internist
Little Rock, AR
University of Arkansas for Medical Sciences
Nominated by the American College of Physicians

Constantine Manthous, MD, Critical care physician
Hamden CT
Yale University
Nominated by the American Thoracic Society

Brenda Snyder, RN, MS, CNS, CCRN, Critical care nurse
Evans, CO
University of Northern Colorado
Nominated by the American Association of Critical-Care Nurses

Mark W. Thomas, RPh, MS, Pharmacist, Pediatrics
Minneapolis, MN
Children´s Hospital and Clinics-Minneapolis, St. Paul
Nominated by the American Society of Health-system Pharmacists

Mark Williams, MD, Hospitalist
Atlanta, GA
Emory University of Medicine
Nominated by the National Association of Inpatient Physicians

Charles Yowler, MD, Surgeon, Critical Care - Burn Surgery
Cleveland, OH
Case Western Reserve University
Nominated by the American College of Surgeons

In 2013, ten panels of experts were convened to support the process of converting the AHRQ QIs from ICD-9-CM to ICD-10-CM/PCS in an accurate and transparent manner, to improve the validity and usefulness of the QIs. Four of these panels –focused on Cancer, Infection, Medicine, and Surgery - advised AHRQ on the ICD-10-CM/PCS specifications for PSI 04. The active members of these panels were:

Ann Borzecki, MD, MPH
Bedford, MA

Dept. of Health Policy and Management, Boston University School of Public Health, and Section of General Internal Medicine, Boston University School of Medicine, and Center for Health Quality, Outcomes and Economic Research Bedford VAMC

B. Ashleigh Guadagnolo, MD, MPH
Houston, TX
The University of Texas MD Anderson Cancer Center

Danil Victor Makarov, MD
New York, NY
Dept of Urology, New York University School of Medicine

Gail Grant, MD, MPH, MBA
Los Angeles, CA
Resource & Outcomes Management, Cedars-Sinai Health System

Joel V. Brill, MD, AGAF
Bethesda, MD
AGA Digestive Health Outcomes Registry
Fair Health, Inc., New York

John Maa, MD
San Francisco, CA
Dept of Surgery, UCSF

Kay Schwebki, MD, MA, MPH
Eden Prairie, MN
OptumInsight

Richard Dutton, MD, MBA
Park Ridge, IL
Anesthesia Quality Institute

Robert S. Gold, MD (expired 2016)
Atlanta, GA
CEO, DCBA, Inc

Coding Professionals:
Bobbi Moore, MBA, RHIT
Grand Rapids, MI
AHIMA Approved ICD-10-CM/PCS Trainer
Quality & Safety Department, Spectrum Health

Gloryanne Bryant, BS, RHIA, RHIT, CCS, CDIP, CCDS
Oakland, CA
AHIMA Approved ICD-10-CM/PCS Trainer
NCAL Revenue Cycle – HIM, Kaiser Foundation Health Plan, Inc.

Jennifer Hornung Garvin, PhD, MBA, RHIA, CPHQ, CCS, CTR
Salt Lake City, UT
FAHIMA, AHIMA Approved ICD-10-CM/PCS Trainer
IDEAS Research Center, VA Salt Lake Health Care System and University of Utah Dept of Biomedical Informatics

Lou Ann Schraffenberger, MBA, RHIA, CCS, CCS-P
Oak Brook, IL
FAHIMA, AHIMA Approved ICD-10-CM/PCS Trainer
Advocate Health Care

Mary Johnson, RHIT, CCS-P
Germantown, OH
Dept of Veteran Affairs

Monica VanSuch, MBA, RHIA
Rochester, MN
Division of Health Care Policy and Research, Mayo Clinic

Nancy Andersen, RHIA, CCS, CRCR,
Oakland, CA
AHIMA Approved ICD-10-CM/PCS Trainer
National Compliance, Ethics, and Integrity Office, Kaiser Foundation Health Plan, Inc.

Rayna Scott, MS, RHIA, CHDA
Oakbrook Terrace, IL
Division of Healthcare Quality Evaluation
The Joint Commission

Sandra Bailey, RHIA
Warrenton, VA
AHIMA Approved ICD-10-CM/PCS Trainer
Cooper Thomas

Sandra Seabold, MBA, RHIA
Cleveland, OH
Cleveland Clinic

Nurses:
Irene Lopez, BSN, RN,CSTR
Austin, TX
Trauma Services Administration, University Medical Center Brackenridge

Karen Snyder, BSN, RN
Cleveland, OH
Cleveland Clinic

Kathleen Hartman, RN, MSN
Cleveland, OH
Cleveland Clinic

Kathryn Fiandt, PhD, RN, FNP-BC, FAANP, FAAN
Galveston, TX
School of Nursing, University of Texas Medical Branch

Marybeth Farquhar, PhD, MSN, RN
Washington, DC
URAC

Patricia Hildebrand, RN, MSN, CCS-P, CPHQ, FACHE
Sugarland, TX
Hildebrand Healthcare Consulting, LLC

Other Professionals:
Anthony Warmuth, MPA, FACHE, CPHQ
Cleveland, OH
Office of Quality, Cleveland Clinic

Catherine Fulton, BS, MS, CPHQ
Montpelier, VT
Vermont Program for Quality in Health Care, Inc

James Notaro, PhD
New York, NY
Clinical Support Services, Inc.

Tina Hernandez-Boussard, PhD, MPH
Palo Alto, CA
Division of General Surgery, Stanford University School of Medicine

Wendy Patterson, MPH
Albany, NY
Office of Quality & Patient Safety, New York State Dept of Health

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2002

Ad.3 **Month and Year of most recent revision:** 06, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 06, 2016

Ad.7 **Disclaimers:** None

Ad.8 **Additional Information/Comments:** None

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 1551
Measure Title: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
Measure Steward: Centers for Medicare & Medicaid Services
Brief Description of Measure: The measure estimates a hospital-level risk-standardized readmission rate (RSRR) following elective primary THA and/or TKA in Medicare Fee-For-Service beneficiaries who are 65 years and older. The outcome (readmission) is defined as unplanned readmission for any cause within 30 days of the discharge date for the index admission (the admission included in the measure cohort). A specified set of planned readmissions do not count in the readmission outcome. The target population is patients 65 and over. CMS annually reports the measure for patients who are 65 years or older, are enrolled in fee-for-service (FFS) Medicare, and hospitalized in non-federal acute-care hospitals.
Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized readmission rates (RSRRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA readmission is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting readmission rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices, as well as increase transparency for consumers. In addition, it has the potential to lower health care costs associated with readmissions.

Numerator Statement: The outcome for this measure is 30-day readmission. We define readmission as an inpatient admission for any cause, with the exception of certain planned readmissions, within 30 days

from the date of discharge of the index hospitalization. If a patient has more than one unplanned admissions (for any reason) within 30 days after discharge from the index admission, only one is counted as a readmission. The measure looks for a dichotomous yes or no outcome of whether each admitted patient has an unplanned readmission within 30 days. However, if the first readmission after discharge is considered planned, any subsequent unplanned readmission is not counted as an outcome for that index admission, because the unplanned readmission could be related to care provided during the intervening planned readmission rather than during the index admission.

Denominator Statement: The target population for the publicly reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures. Additional details are provided in S.9 Denominator Details.

Denominator Exclusions: This measure excludes admissions for patients:

1) Without at least 30 days post-discharge enrollment in FFS Medicare;
2) Who were discharged against medical advice (AMA);
3) Admitted for the index procedure and subsequently transferred to another acute care facility;
4) Who had more than two THA/TKA procedure codes during the index hospitalization; or
5) Who had THA/TKA admissions within 30 days of a prior THA/TKA index admission.

Measure Type: Outcome
Data Source: Administrative claims, Other
Level of Analysis: Facility

Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 31, 2012

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This outcome measure was initially endorsed in 2012 and calculates risk-standardized readmission rates following elective total hip arthroplasty (THA) and/or total knee arthroplasty (TKA).
- The developer provided information and a diagram illustrating the relationship between timely and high quality care, improved communication between providers, and patient education and decreased risk of readmission.
- New evidence submitted discusses the number of THA and TKA procedures performed in 2010.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *The developer attests the underlying evidence for the measure has not changed since the last NQF endorsement review. Does the Committee agree the evidence basis for the measure has not changed and there is no need for repeat vote on Evidence?*

Guidance from the Evidence Algorithm: Health outcome (Box 1) →relationship between outcome and at least one healthcare action identified/supported by stated rationale (Box 2) → Pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Data submitted during the previous review show variation in rates across hospitals. The unadjusted mean readmission rate was 6.78% and ranged from 0 to 100% across 3,310 hospitals in 2008. After adjustment for patient and clinical factors, the mean readmission rate was 6.30%, ranging from 3.06% to 50.94%.
- Current performance data were analyzed from over 3,000 hospitals between 2011 and 2014. The median and mean risk standardized readmission rate (RSRR) for all measured hospitals in the most recent reporting period (07/2011 through 06/2014) was 4.8% and 4.9%, respectively.

	07/2011-06/2012	07/2012-06/2013	07/2013-06/2014	07/2011-06/2014
Number of hospitals	3,348	3,331	3,307	3,498
Number of admissions	302,352	306,937	317,396	926,685
Mean (SD)	5.3 (0.4)	4.9 (0.5)	4.4 (0.4)	**4.9 (0.5)**
Range	3.6-7.6	2.9-7.5	2.8-6.4	2.6-8.6
50th percentile	5.3	4.8	4.4	**4.8**

Disparities

- Disparities data show that risk-standardized readmission rates are similar among patients with social risk factors compared to patients without these risk factors. The developer reports RSRR rates tend to be slightly higher among hospitals with higher proportions of patients with these risk factors. The developer notes that the difference in median readmission rates ranges from 0.0 to 0.2 absolute percentage points depending upon the social risk factor measure, indicating relatively small differences across groups.
- Note that data for dual eligible and African-American patients come from Medicare FFS claims; data for patients with an AHRQ SES index score below 42.7 come from Medicare FFS claims and the American Community Survey (2009-2013). AHRQ SES index scores describe the socioeconomic status of people living in defined geographic areas.

Distribution of THA/TKA RSRRs by Characteristic, July 2013-June 2014

	Proportion of Dual Eligible		Proportion of African American Patients		Proportion of Patients with AHRQ SES Index Scores <42.7	
	Hospitals with Low Proportion	Hospitals with High Proportion	Hospitals with Low Proportion	Hospitals with High Proportion	Hospitals with Low Proportion	Hospitals with High Proportion
% of hospitals with characteristic	3.9%	11.8%	0.0%	6.3%	6.4%	24.0%
Number of Patients	329,366	107,228	96,689	228,821	274,914	138,643
Median	4.7	4.9	4.8	5.0	4.7	4.9

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support the Measure Focus

- Again, why just old people?
  Patients transferred to another facility are excluded. How does this affect the results? does it allow gamesmanship? are rehab hospitals "another facility?"
  Very low volume facilities excluded. Is this a group that should be targeted, or does this allow a 'learning curve?'
  49 better and 49 worse out of >3,000 facilities
- This is an OUTCOME measure and calculates risk-standardized readmission rates after joint replacement (hip or knee) of Medicare beneficiaries (> 65 years old or dual-eligible patients).
- Yes. The evidence shows small differences for events with a low incidence. Question: Is the contribution of statistical noise adequately separable when comparing individual hospitals on a continuous value as opposed to reporting of classes of hospitals (below, expected, above performance)? The question is made in the context of the measure now being collected for application to the Readmission Reduction program.

1b. Performance Gap

- There is wide geographic variation in surgical rate of THA and TKA across the country.
  The unadjusted 30 day readmission rate was 6.3% (range 3.06% to 50.94%)
  The mean hospital performance score was 4.9 (range 2.6 to 8.6).
- The performance gap is the variability in rates of readmissions. The variability is adequate.
  In terms of potential disparities, there were more readmissions in the hospitals with higher proportions of African-American patients, those with dual eligibility, and those patients with AHRQ SES scores less than 42.7. These results were discounted as being more affected by the treating hospital than the patients.
  My question for the stewards is based on the following. If one accepts that the race in and of itself does not add risk, the race distinction can be, unfortunately in our society, a surrogate for poverty, in particular in urban areas. Just because the other patients at that hospital are not

African Americans, might not have dual eligibility, and might be just above the cut-off for the SES designation, it does not mean that they are "rich". In fact, they are more likely to be living in varying degrees of poverty. The neighborhoods that such a hospital cares for is overall more likely to be caring for a spectrum of poverty, that if averaged across all patients, would score a lower SES than hospitals with lower proportions of such patient classes.

Social and economic dysfunction of the poorest communities would ordinarily correlate with less health care sophistication, poorer life choices, higher degrees of alcohol and drug use, less access to care and transportation, and living environments that are not ideal. This would be the common denominator across the community at large from in which the population subsets of minorities, dual eligibility, and those patients with SES scores less than 42.7 live.

If the overall composition of the hospital's population is not taken into consideration as a risk factor for performance, is the hospital effect being "double counted" in a negative way? This is a testable question; if the above argument is true, the distribution of SES scores across the entirety of the hospital population should be lower in hospitals with higher proportion of AA, DE, and the SES cut-off score. It is the relative poverty of the entire patient population, not just those that can be defined by dichotomous factors including those patients with an SES score below 42.7. In effect, the very best hospital in an impoverished urban environment might be encumbered with social obstacles affecting all patients in ways that a suburban hospital does not encounters; it is the community served as a whole, not just the selected risk factors, that could need risk adjustment and provide disparities.

Criteria 2: Scientific Acceptability of Measure Properties
2a. Reliability
2a1. Reliability Specifications **Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures**
2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented. **Data source(s):** The developer lists administrative claims and Census Data/American Community survey as data sources. **Specifications:** • The measure is specified as a facility level measure for the hospital/acute care setting. • The numerator includes readmissions to any acute care hospital for any cause within 30 days of the date of discharge of the index THA/TKA hospitalization. • The denominator includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA/TKA procedures. • ICD-9 and ICD-10 codes and a crosswalk table are included in the data dictionary • Exclusions listed are patients without at least 30 days post-discharge enrollment in FFS Medicare; patients discharged against medical advice; patients transferred to another facility; patients with more than two THA/TKA procedure codes during the index hospitalization; and patients who had THA/TKA admissions within 30 days of a prior THA/TKA admission.

- The developer also notes that planned readmissions are not considered readmissions in the measure outcome.
- This outcome measure is risk-adjusted using a statistical risk-model with 33 risk factors.
- The measure is calculated as the ratio of the number of predicted to the number of expected readmissions at a given hospital, multiplied by the national observed readmission rate.
- The developer notes that changes to the specifications since it was last endorsed include updating to the CMS Planned Readmission Algorithm as changes were made (including removing two procedure categories and adding several acute diagnoses), and exclusion of patients with secondary diagnosis of fracture during index admission. A summary of all measure specification updates are here.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is there data/information that the Committee needs with respect to changes to the specifications since last endorsement?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- The analysis submitted for reliability testing for the previous evaluation demonstrated similar risk model performance using development and validation samples. This testing does not meet current NQF requirements for reliability testing.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Dataset 1 used for testing included Medicare Part A and Part B, and the Medicare Enrollment database for the 2015 reporting cohort (from July 1, 2011 – June 30, 2014). The dataset included 926,685 admissions from 3,498 hospitals.
- Developers used a split-sample methodology to test Measure score reliability. The dataset was randomly split into two samples; 50% of Medicare patients 65 and over in the most recent 3-year cohort and the remaining 50% comprised the second group. The level of agreement between scores was compared using the intra-class-correlation coefficient (ICC). The ICC demonstrates the percentage of variance in score results that is due to true or real variance between hospitals.
- Although the developer reported assessing data element reliability by comparing model variable frequencies and odds ratios from logistic regression models across the most recent three years of data, NQF does not consider temporal consistency to be a valid method of demonstrating reliability of data elements.

Results of reliability testing

- Measure score reliability results
  - The dataset included 926,685 admissions from 3,498 hospitals, with 460,576 index admissions from 2,835 hospitals in one sample and 459,237 admissions from 2,835 hospitals in the other. The ICC was 0.49, indicating that 49% of the variance in scores are due to differences between hospitals. According to the Landis and Koch classification, an ICC value of 49% can be interpreted as moderate agreement. However, a value of 0.7 is often regarded as a minimum acceptable reliability value.
  - The developer notes that the analysis is limited to hospitals with 12 or more cases in each split sample. The measure is not specified to include a minimum data sample of 12 cases. The ICC is based on a split sample of three years data, resulting in a volume of patients in each sample to 1.5 years of data. The measure is reported with the full three years of data.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is a minimum data sample limitation to hospitals with 12 or more cases appropriate for this measure?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) →empiric reliability testing (Box 2) →performance score testing (Box 4) →appropriate method of testing (Box 5) → moderate certainty of reliability (Box 6b)

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:
SUMMARY OF TESTING
Validity testing level ☒ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer states that measure validity is demonstrated through prior validity testing on its claims-based measures, through use of established measure development guidelines, and by

systematic assessment of measure face validity by a technical advisory panel. The processes are described. The precise methodology and results are not provided. NQF does not consider technical panel review an appropriate method of demonstrating face validity.

- The developer also reports validation efforts for two procedure based complications measures, one of which is elective primary THA/TKA, noting that there was strong agreement between complications coded in claims and abstracted medical record data but did not provide detail regarding methods or results.

Validity testing results:

- The TEP assessed the face validity of the measure and found the measure valid.
- Information provided suggests that measure testing may have been at the data element, rather than measure score, level.

Questions for the Committee:

- *Does the information about testing as well as the information about changes to the measure over time give sufficient information for the Committee to opine on the validity of the measure?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

- To determine the impact of exclusions on the cohort, the developer examined overall frequencies and proportions of the total 2015 cohort excluded for each exclusion criterion.
  - Patients discharged against medical advice -0.01%
  - Patients without at least 30 days post-discharge enrollment in FFS Medicare -0.20%
  - Patients admitted for the index procedure and transferred to another acute care facility – 0.91%
  - Patients who had more than two THA/TKA procedure codes – 0.0%
  - Patients who had an admission for THA/TKA within 30 days of a prior index admission – 0.15%
- The developer stated that exclusions for patients discharged against medical advice and those without at least 30 days post discharge enrollment were not likely to affect the measure score since a small percentage of patients were excluded. The remaining exclusions were needed to ensure patients are accurately counted in the measure.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment method: ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☒ **Yes** ☐ **No**
SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary

Description of the model

- This measure is risk adjusted using the hierarchical logistic regression model with 33 factors to create a hospital-level 30 days RSRR. This approach accounts for variance in outcomes within and

between hospitals.
    - The model adjusts for age (65 and older), male gender, index admissions with an elective THA and/or TKA procedure, number of procedures performed, and several clinical risk factors.
    - Only co-morbidities that conveyed information about the patient at that time or in the 12 months prior, and not in complication that arose during the course of the admission, were included in the risk adjustment.
- Candidate variables were patient level risk adjustors that were expected to be predictive of readmission, based on empirical analysis, prior literature, and clinical judgement. For each patient, covariates were obtained from Medicare claims extending 12 months prior to and including the index admission. The model adjusts for case differences based on the clinical status of the patient at the time of the admission using condition categories.

Performance of the model

Discrimination statistics:

- The c-statistic reflects how accurately a statistical model distinguishes between a patient with and without an outcome. C-statistic values range from 0.5 to 1.0, where a value of 0.5 indicates the model is no better than chance at making a prediction of patients with and without the outcome of interest.
- Dataset 1 is the 2015 cohort of Medicare Parts A and B data, and the Medicare Enrollment Database. Dataset 2 is calendar year 2008 Medicare Parts A and B.
    - The c-statistic for Dataset 2 development sample was 0.65 (lowest decile 2.4%, highest decile, 13.4%)
    - The c-statistic for the Dataset 2 validation sample was 0.64 (lowest decile 2.6%, highest decile 13.2%)
    - The c-statistic for Dataset 1 (current measure cohort) was 0.65 (lowest decile 1.8%, highest decile 10.9%).
- The developer notes that c-statistics for Dataset 1 and Dataset 2 indicate consistent and fair model discrimination. The models indicated a wide range between the lower and highest decile, indicating the ability to distinguish between high risk and low risk patients.

Calibration statistics:

- Calibration statistics for Dataset 2 demonstrate good calibration of the model with values close to 0 at one end and close to 1 at the other end for both of the two split samples.

Risk Decile Plot:

- The developer notes that the risk decile plots for Dataset 1 demonstrated excellent discrimination of the model and good predictive ability since higher deciles of the predicted outcomes are associated with higher observed outcomes.

Overall

- Interpreting the three diagnostic results together, the developer states that the risk adjustment model adequately controls for differences in patient characteristics.

Conceptual basis and empirical support for potential inclusion of SDS factors in risk-adjustment approach

- The developer completed a literature review that found that SES and race variables may be associated with increased risk of THA/TKA readmission. The developer cites additional studies that have found significant differences in the rate of THA received by African-American and white patients, indicating that patient and surgeon behavior could also contribute to disparities based on racial factors. The developers note that overall there is no clear consensus on which risk factors demonstrate the strongest relationship with readmission.
- The developer identified 4 conceptual pathways to consider:
    - Relationship of SES factors or race to health at admission

- Use of low quality hospitals – The developer states that patients of lower income, lower education or unstable housing have been shown to hot have access to high quality facilities since these facilities are less likely to located in areas with large populations of poor patients
- Differential care within a hospital – The developer states that African-American patients may experience differential, lower quality or discriminatory care; patients of lower education may also require differentiated care that clients may not necessarily receive
- Influence of SES on readmission risk outside of hospital quality and health status

- Based on the interpretation of the literature and analysis of the conceptual pathways, three SES and race variables were considered (listed below). Analyses of the strength and significance of the SES and race variables in a multivariable model found the effect size of each of the variables to be moderate, with the c-statistic nearly unchanged. Additionally, the inclusion of these variables in the model had little to no effect on hospital performance.
  - Factor – Effect Size, Median absolute change in hospital RSRR
    - Dual-eligible status – OR 1.22, 0.0199%
    - African-American race – OR 1.19, 0.0217%
    - AHRQ SES Index – OR 1.09, 0.0172%
- Decomposition analysis found that all three factors were significantly associated with THA/TKA readmission. The developer states that if these variables were used in the model to adjust for patient level differences, then some of the differences between hospitals would also be adjusted which could hide signs of hospital quality.
- Based on these results the developer decided not to include any of the SDS factors analyzed in the final risk-adjustment model.

Questions for the Committee:

- *Is an appropriate risk-adjustment strategy included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's decision, based on their analysis, to not include SDS factors in their risk-adjustment model?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- Analyses of Medicare FFS Dataset 1 (July 2011 – June 2014), showed variation in RSRR among hospitals, with the median RSRR at 4.8%, ranging from 2.6% to 8.6%. The interquartile range was 4.6% to 5.2%. The developer notes that the variation in rates and number of outliers (49 hospitals performed better and another 49 performed worse than the US national rate) suggest that differences remain in the quality of care received across hospitals for THA/TKA.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- Not applicable

2b7. Missing Data

- The developer did not provide an analysis for missing data, although they note in S.22 that "*missing values are rare among variables used from claims data in this measure*".

Guidance from Validity Algorithm: Precise specifications (Box 1) → potential threats to validity assessed (Box 2) → empirical validity testing (Box 3) →measure score testing (Box 4) → Insufficient

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

- This measure targets the facility level (hospital).
  Numerator and denominator are clearly defined.
  It is evident in the evidence provided that CMS is trying to keep up with the evolving science of delineating an unplanned 30 day readmission from a planned 30 day readmission now utilizing the CMS Planned Readmission Algorithm (Version 4.0) with updates from the 2015 version of the AHRQ CCS.
- The data elements are expected to be reliable. Recent changes adding more excluded readmissions improve the model.

2a2. Reliability – Testing

- Reliability was determined by testing two independent data sets (data set 1 & data set 2). Using this split sample method, the developers tested the measure score reliability (not the data element reliability). Data element reliability was not assessed.

  ICC was 0.49. This is can be interpreted as moderate agreement or a value of 0.7 can be regarded as a threshold value for reliability

- The ICC is 49% (moderate)
  The c-statistic is 0.65.

2b.1 Validity – Specifications

- This is an outcome measure, not a process measure.
  Face validity was from a TEP panel.
  A prior NQF surgical committee accepted this as valid (2012).

  The target population is Medicare patients over the age of 65 years undergoing joint replacement in the Medicare FFS model.
  Other medicare beneficiaries who are excluded would include Medicare HMO patients and dual-eligible patients under the age of 65.

- The specifications are consistent with the evidence.

  The original endorsement was found to be valid by it's TEP and the NQF for public reporting using three categories of below expected, expected, and above expected levels of performance.

  It is now being used for assigning penalties in the Readmission Reduction Program. It is being used as a continuous value. Does the face validity of ascribed to the measure from the TEP apply to these new uses?

2b2. Validity – Testing

- Face validity was from a TEP panel.
  A prior NQF surgical committee accepted this as valid (2012).

  The developers write, "We have also completed two national, multi-site validation efforts for two procedure-based complication measures (elective primary THA/TKA and implantable cardioverter defibrillator). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical record data."

  Can the developers please provide the citation for this work?

- The ability to determine readmission is more defined than the capturing of the complications in 1550. No specific validation study for this measure was performed, but validation of the ability to capture the risk factors used has been presented/published. The measure has, in previous forms, used the validation study of the complication agreement between administrative data and charts as a surrogate source of proving face validity. Please see the discussion for NQF 1550 for questions about that study. Given the capture of the risk factors over time, the older validity studies used are probably more appropriate and acceptable.

2b3-7. Threats to Validity

- This measure is limited by the limits of Medicare FFS claims data:
  1) Restricted to adults over the age of 65 in Medicare FFS
  2) Dual-eligibles under the age of 65 are excluded
  3) Claims data is a proxy, but not as reliable as chart-abstraction

  The exclusion criteria with rationale (AMA, without 30-days of FFS Medicare, transferred to another facility, more than 2 THA/TKA, prior index admission in 30 days previous) are very clear.

  Can the developers clarify if they use a current medical comorbidity index based on claims data (e.g. the Elixhauser co-morbidity index is updated annually to account for changes in billing codes)?

  The c-statistic of the risk adjusted models in the development and validation cohorts were 0.65 and 0.64.

  The developer considered three SES variables:
  1) Dual-eligible status (OR 1.22, 0.0199%)
  2) African American race (OR 1.19, 0.0217%)
  3) AHRQ SES index (OR 1.09, 0.0172%)
  And decided not to include any of the SDS factors in the final model.

- please see questions regarding SDS in the evidence section above

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- All data elements are in defined fields in electronic claims and generated or collected by and used by health care personnel during the provision of care. The data are coded by someone other than the person obtaining original information.
- Administrative data are routinely collected as part of the billing process.
- There are no fees associated with the use of the measure.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Preliminary rating for feasibility: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- Because this measure is based on claims data alone, feasibility is good.
- Feasible as designed.

  Adding other risk factors to the risk model might need lower the feasibility due to reporting burdens and capture rates

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☐ **Yes** ☐ **No**

Accountability program details

- This measure is currently used in the CMS Hospital Inpatient Quality Reporting Program and the Hospital Readmission Reduction (HRRP) Program.

Improvement results

- The developer notes progress in 30-day RSRR for THA/TKA. The median 30-day RSRR decreased by 0.8 absolute percentage points from July 2011 to June 2012 (median RSRR: 5.2%) to July 2013-June 2014

(median RSRR: 4.4%).

Unexpected findings (positive or negative) during implementation

- The developer notes there are no unexpected findings to report.

Potential harms

The developer states that the benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

Feedback : N/A

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- This is a publicly reported measure.
- It is included in HospitalCompare.

  It is now being used as part of the condition basket in the Readmission Reduction Program that can lead to up to a 3% of CMS payments penalty. The rate created is used as a continuous value in the calculation and the components of the basket are weighted. This program represents a zero sum competition over a low value with a relatively small standard deviation. Risk shedding is a potential issue, especially given the low c-statisitc. Please see the discussion on this question in the reply to NQF1550. The concern is that avoidance of perceived to be higher risk condition classes is probable given the leverage of the size of the penalty and assumption of similar behavior by other hospitals.

Criterion 5: Related and Competing Measures

Related or competing measures

- 0330: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization
- 0505: Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following acute myocardial infarction (AMI) hospitalization.
- 0506: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following pneumonia hospitalization
- 1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
- 1789: Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)

- 1891: Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization

Harmonization

- The developer notes the measure is harmonized with all related measures.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1551
Measure Title: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: N/A

Date of Submission: 5/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the

measured process leads to a desired health outcome.

- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Health outcome: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

☐ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process: Click here to name the process

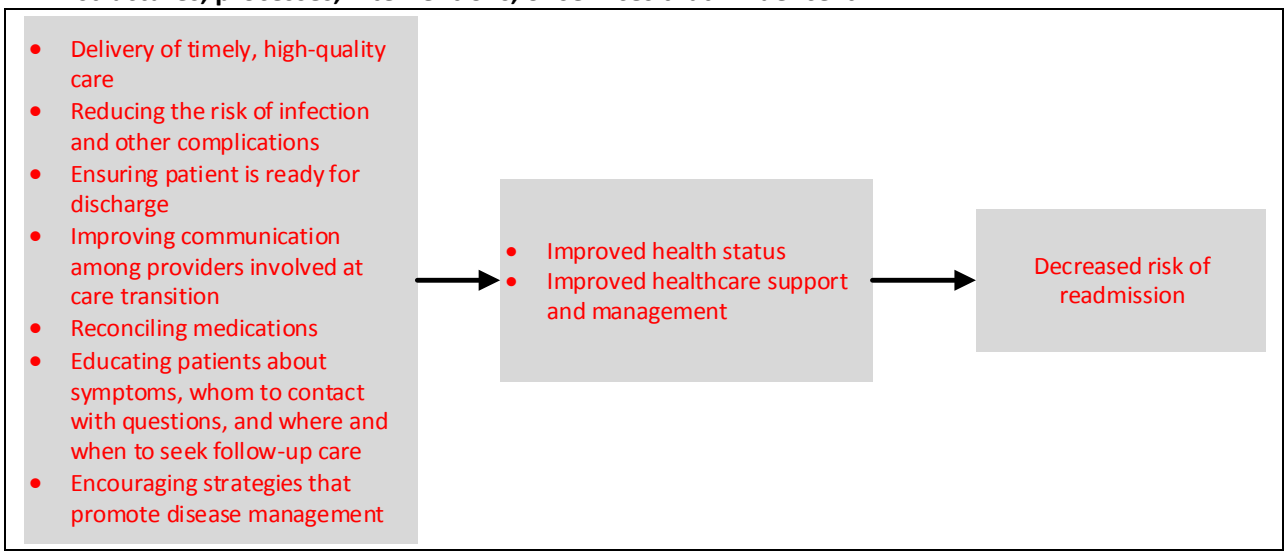
☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.



The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized readmission rates following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). Measurement of patient outcomes allows for a broader view of a hospital's quality of care that encompasses more than what can be captured by individual process of care measures. More specifically, complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This readmission measure was developed to identify institutions, whose performance is better or worse than expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about the quality of care.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

In 2010 there were 168,000 THAs and 385,000 TKAs performed on Medicare beneficiaries 65 years and older (National Center for Health Statistics, 2010). Although these procedures dramatically improve quality of life, they are costly. In 2005, annual hospital charges totaled $3.95 billion and $7.42 billion for primary THA and TKA, respectively (Kurtz et al., 2007). These costs are projected to increase by 340% to 17.4 billion for THA and by 450% to 40.8 billion for TKA by 2015 (Kurtz et al., 2007). Medicare is the single largest payer for these procedures, covering approximately two-thirds of all THAs and TKAs performed in the US (Ong et al., 2006). Combined, THA and TKA procedures account for the largest procedural cost in the Medicare budget (Bozic et al., 2008).

Measuring and reporting elective primary THA/TKA readmission rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and promote improvements in the quality of care received by Medicare patients and the outcomes they experience. The measure will also provide patients with information that could guide their choices regarding where they seek care for these elective procedures. Furthermore, the measure will increase transparency for consumers and has the potential to lower health care costs by reducing the risk of readmissions.

The THA/TKA hospital-specific risk-standardized readmission rate (RSRR) measure is thus intended to inform quality-of-care improvement efforts, as individual process-based performance measures cannot encompass all the complex and critical aspects of care within a hospital that contribute to patient outcomes.

References:

Bozic KJ, Rubash HE, Sculco TP, Berry DJ. An analysis of medicare payment policy for total joint arthroplasty. *J Arthroplasty.* Sep 2008;23(6 Suppl 1):133-138.

Kurtz SM, Ong KL, Schmier J, et al. Future clinical and economic impact of revision total hip and knee arthroplasty. J Bone Joint Surg Am. Oct 2007;89 Suppl 3:144-151.

National Center for Health Statistics. National Hospital Discharge Survey: 2010 table, Procedures by selected patient characteristics - Number by procedure category and age. Available at http://www.cdc.gov/nchs/data/nhds/4procedures/2010pro4_numberprocedureage.pdf.

Ong KL, Mowat FS, Chan N, Lau E, Halpern MT, Kurtz SM. Economic burden of revision hip and knee arthroplasty in Medicare enrollees. Clin Orthop Relat Res. May 2006;446:22-28.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

N/A

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

N/A

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

N/A

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

N/A

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***
☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

N/A

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

N/A

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

N/A

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

N/A

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

N/A

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

N/A

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

N/A

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE
If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

N/A

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

N/A

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

N/A

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

N/A

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

N/A

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

N/A

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

N/A

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

N/A

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

N/A

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

N/A

1a.8.2. Provide the citation and summary for each piece of evidence.

N/A

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

NQF_1551_HipKnee_Readmission_NQF_Measure_Evidence_Form_v1.0.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*

The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized readmission rates (RSRRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA readmission is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting readmission rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices, as well as increase transparency for consumers. In addition, it has the potential to lower health care costs associated with readmissions.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of**

analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
We examine the distribution of hospital performance scores to demonstrate the current gap in quality among measured hospitals. The results below indicate that the median RSRR for all measured hospitals in the most recent 3-year reporting period was 4.8. The mean was 4.9 and the range was 2.6 to 8.6 showing persistent variation in readmission rates across hospitals.

Distribution of Hospital THA/TKA RSRRs over Different Time Periods
Results for each data year
Characteristic//07/2011-06/2012//07/2012-06/2013//07/2013-06/2014//07-2011-06/2014
Number of Hospitals/3,348/ /3,331/ /3,307/ /3,498
Number of Admissions/302,352/ /306,937/ /317,396/ /926,685
Mean (SD)/5.3 (0.4)//4.9 (0.5)/ /4.4 (0.4)/ /4.9 (0.5)
Range (min. – max.)/3.6-7.6/ /2.9-7.5/ /2.8-6.4/ /2.6-8.6
Minimum/3.6/ /2.9/ / 2.8 / /2.6
10th percentile/ 4.8/ / 4.4 / /4.1/ /4.2
20th percentile/ 5.0/ /4.6 / /4.2/ /4.5
30th percentile/ 5.1/ /4.7 / /4.3/ /4.6
40th percentile/5.2 / /4.8 //4.4 //4.8
50th percentile/5.3/ /4.8/ /4.4/ /4.8
60th percentile/5.3/ /4.9 / /4.5/ /4.9
70th percentile/5.4/ /5.0 / /4.6/ /5.1
80th percentile/5.6/ /5.1 / /4.7/ /5.3
90th percentile/5.8/ /5.4 / /4.9/ /5.6
Maximum/7.8/ /7.5/ /6.4/ /8.6

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
The results of this analysis show that risk-standardized readmission rates are similar among patients with social risk factors compared to patients without these risk factors, although they tend to be slightly higher among hospitals with higher proportions of patients with these risk factors. However, the difference in median readmission rates ranges from 0.0 to 0.2 absolute percentage points depending upon the social risk factor measure, indicating relatively small differences across groups.

Distribution of THA/TKA RSRRs by Proportion of Dual Eligible Patients:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims
Characteristic//Hospitals with a low proportion (=3.9%) Dual Eligible patients//Hospitals with a high proportion (=11.8%) Dual Eligible patients
Number of Measured Hospitals// 704 // 704
Number of Patients// 329,366 patients in low-proportion hospitals// 107,228 patients in high-proportion hospitals
Maximum// 7.3// 8.6
90th percentile// 5.5// 5.7

75th percentile// 5.1// 5.3
Median (50th percentile)// 4.7// 4.9
25th percentile// 4.3// 4.6
10th percentile// 4.0// 4.3
Minimum // 3.1// 3.4

Distribution of THA/TKA RSRRs by Proportion of African-American Patients:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims
Characteristic// Hospitals with a low proportion (=0.0%) African-American patients//Hospitals with a high proportion (=6.3%) African-American patients
Number of Measured Hospitals// 715// 704
Number of Patients// 96,689 patients in low-proportion hospitals// 228,821 patients in high-proportion hospitals
Maximum// 6.5// 8.6
90th percentile// 5.4// 6.0
75th percentile// 5.1// 5.5
Median (50th percentile)// 4.8// 5.0
25th percentile// 4.5// 4.7
10th percentile// 4.2// 4.3
Minimum // 3.4// 3.4

Distribution of THA/TKA RSRRs by Proportion of Patients with AHRQ SES Index Scores Below 42.7:
Dates of Data: July 2013 through June 2014 Data Source: Medicare FFS claims and the American Community Survey (2009-2013) data
Characteristic//Hospitals with a low proportion of patients below AHRQ SES index score of 42.7 (=6.4%)// Hospitals with a high proportion of patients below AHRQ SES index score of 42.7 (=24.0%)
Number of Measures Hospitals// 706// 705
Number of Patients// 274,914 patients in hospitals with low proportion of patients below AHRQ SES index score of 42.7// 138,643 patients in hospitals with high proportion of patients below AHRQ SES index score of 42.7
Maximum// 7.3// 7.6
90th percentile// 5.6// 5.8
75th percentile// 5.1// 5.3
Median (50th percentile)// 4.7// 4.9
25th percentile// 4.4// 4.6
10th percentile// 4.0// 4.3
Minimum // 2.6// 3.4

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF;
  OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

THA and TKA are commonly performed procedures that improve quality of life. In 2003, there were 202,500 THAs and 402,100 TKAs performed (Kurtz et al., 2007a) and the number of procedures performed has increased steadily over the past decade (Kurtz et al., 2007b; Ong et al., 2006).

Although these procedures dramatically improve quality of life, they are costly. In 2005, annual hospital charges totaled $3.95 billion and $7.42 billion for primary THA and TKA, respectively (Kurtz et al., 2007b). These costs are projected to increase by 340% to $17.4 billion for THA and by 450% to $40.8 billion for TKA by 2015 (Kurtz et al., 2007b). Medicare is the single largest payer for these procedures, covering approximately two-thirds of all THAs and TKAs performed in the US (Ong et al., 2006). Combined, THA and TKA procedures account for the largest procedural cost in the Medicare budget (Bozic et al., 2008).

Hospital readmission is an outcome that is influenced by quality of care and is an important outcome for patients. Hospital processes that reflect the quality of inpatient and outpatient care such as discharge planning, medication reconciliation, and coordination of outpatient care have been shown to reduce readmission rates (Nelson, Maurish, & Axler, 2000). Although readmission rates are also influenced by hospital system characteristics, such as the bed capacity of the local health care system (Fisher et al., 1994), these hospital characteristics should not influence quality of care. Therefore, this measure does not risk-adjust for such hospital characteristics.

Measuring and reporting elective primary THA/TKA readmission rates will inform health care providers about opportunities to improve care, strengthen incentives for quality improvement, and promote improvements in the quality of care received by Medicare patients and the outcomes they experience. The measure will also provide patients with information that could guide their choices regarding where they seek care for these elective procedures. Furthermore, the measure will increase transparency for consumers and has the potential to lower health care costs by reducing the risk of readmissions.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Bozic KJ, Rubash HE, Sculco TP, Berry DJ. An analysis of medicare payment policy for total joint arthroplasty. J Arthroplasty. Sep 2008;23(6 Suppl 1):133-138.

Fisher ES, Wennberg JE, Stukel TA, Sharp SM. Hospital Readmission Rates for Cohorts of Medicare Beneficiaries in Boston and New Haven. New England Journal of Medicine. 1994;331(15):989-995.

Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am. Apr 2007;89(4):780-785.

Kurtz SM, Ong KL, Schmier J, et al. Future clinical and economic impact of revision total hip and knee arthroplasty. J Bone Joint Surg Am. Oct 2007;89 Suppl 3:144-151.

Nelson EA, Maruish ME, Axler JL. Effects of Discharge Planning and Compliance With Outpatient Appointments on Readmission Rates. Psychiatr Serv. July 1 2000;51(7):885-889.

Ong KL, Mowat FS, Chan N, Lau E, Halpern MT, Kurtz SM. Economic burden of revision hip and knee arthroplasty in Medicare enrollees. Clin Orthop Relat Res. May 2006;446:22-28.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.**

(Describe how and from whom their input was obtained.)
N/A. This measure is not a PRO-PM.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Musculoskeletal : Joint Surgery, Musculoskeletal : Osteoarthritis, Musculoskeletal : Rheumatoid Arthritis, Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Care Coordination, Care Coordination : Readmissions, Safety, Safety : Complications, Safety : Healthcare Associated Infections, Safety : Readmissions

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.qualitynet.org/dcs/ContentServer?c=Page&pagename=QnetPublic%2FPage%2FQnetTier3&cid=1219069855273
http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.html

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** NQF_1551_HipKnee_Readmission_S2b_Data_Dictionary_v1.0.xlsx

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
Updates by Year

2016
1. Respecified the measure by updating to CMS Planned Readmission Algorithm (Version 4.0).
Rationale: Version 4.0 incorporates improvements made following a validation study of the algorithm using data from a medical record review and input from clinical experts. These changes improve the accuracy of the algorithm by decreasing the number of readmissions that the algorithm mistakenly designates as planned or as unplanned by removing five procedure categories from and adding one procedure category to the list of potentially planned procedures.

2. Applied the 2015 version of the AHRQ CCS to the planned readmission algorithm.
Rationale: A 2015 version of the AHRQ CCS was released. We identified any changes from the previous version that might affect the planned readmission algorithm and, therefore, the assessment of the measure outcome

2015
1. Applied updated AHRQ CCS version to the planned readmission algorithm.
Rationale: An updated version of the AHRQ CCS was released in 2014.

2014
1. Respecified the measure by adding the CMS Planned Readmission Algorithm (Version 3.0).
Rationale: Version 3.0 incorporates improvements made following a validation study of the algorithm using data from a medical record review. These changes improve the accuracy of the algorithm by decreasing the number of readmissions that the algorithm mistakenly designated as planned by removing two procedure categories and adding several acute diagnoses.

2. Updated measure specifications to not include all patients with a secondary diagnosis of fracture during index admission in the measure cohort.
Rationale: These procedures are presumably not elective THA/TKA procedures, and the cohort aims to include only elective THA/TKA procedures.

3. Applied updated AHRQ CCS version to the planned readmission algorithm.
Rationale: An updated version of the AHRQ CCS was released in 2013.

2013
1. Respecified the measure by adding a planned readmission algorithm.
Rationale: Unplanned readmissions are acute clinical events a patient experiences that require urgent rehospitalization. In contrast, planned readmissions are generally not a signal of quality of care. Including planned readmissions in a readmission measure could create a disincentive to provide appropriate care to patients scheduled for elective or necessary procedures within 30 days of discharge.

2. Updated CC map.
Rationale: Prior to 2014, the ICD-9-CM Hierarchical Condition Category (HCC) map was updated annually to capture all relevant comorbidities coded in patient administrative claims data.

3. Changes from prior methodology report.
Rationale: Rationale: There were two changes from the original methodology report.
i. Table A3 contains the updated listing of the ICD-9-CM codes for fractures, malignant neoplasms, revisions, and other procedures that exclude patients from the measure cohort.
ii. The mean risk-standardized readmission rate for the 2008 sample on page 54 was corrected.

4. Updated planned readmission algorithm handling of admissions to psychiatric and rehabilitation hospitals.
Rationale: Psychiatric and rehabilitation hospitals in Maryland have the same provider ID number as acute care hospitals. Therefore, readmissions are not counted if the patient has a principal diagnosis code beginning with a "V57" (indication of admission to a rehab unit) or if all three of the following criteria are met: (1) the admission being evaluated as a potential readmission has a psychiatric principal discharge diagnosis code (ICD-9 codes 290-319); (2) the index admission has a discharge disposition code to a psychiatric hospital or psychiatric unit from the index admission; and (3) the admission being evaluated as a potential readmission occurred during the same day as or the day following the index discharge. The criteria for identifying such admissions are available in the 2010 AMI, HF, and pneumonia readmission measures maintenance report.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The outcome for this measure is 30-day readmission. We define readmission as an inpatient admission for any cause, with the

exception of certain planned readmissions, within 30 days from the date of discharge of the index hospitalization. If a patient has more than one unplanned admissions (for any reason) within 30 days after discharge from the index admission, only one is counted as a readmission. The measure looks for a dichotomous yes or no outcome of whether each admitted patient has an unplanned readmission within 30 days. However, if the first readmission after discharge is considered planned, any subsequent unplanned readmission is not counted as an outcome for that index admission, because the unplanned readmission could be related to care provided during the intervening planned readmission rather than during the index admission.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Numerator Time Window: We define the time period for readmission as within 30 days from the date of discharge of the index THA and/or TKA hospitalization.

Denominator Time Window: This measure was developed with 12 months of data. The time window can be specified from one to three years. Currently, the measure is publicly reported with three years of index admissions.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
The measure counts readmissions to any acute care hospital for any cause within 30 days of the date of discharge of the index THA and/or TKA hospitalization, excluding planned readmissions as defined below.

Planned Readmission Algorithm (Version 4.0)
The Planned Readmission Algorithm is a set of criteria for classifying readmissions as planned among the general Medicare population using Medicare administrative claims data. The algorithm identifies admissions that are typically planned and may occur within 30 days of discharge from the hospital.

The Planned Readmission Algorithm has three fundamental principles:

1. A few specific, limited types of care are always considered planned (transplant surgery, maintenance chemotherapy/immunotherapy, rehabilitation);
2. Otherwise, a planned readmission is defined as a non-acute readmission for a scheduled procedure; and
3. Admissions for acute illness or for complications of care are never planned.

The algorithm was developed in 2011 as part of the Hospital-Wide Readmission measure. In 2013, CMS applied the algorithm to its other readmission measures. In applying the algorithm to condition- and procedure-specific measures, teams of clinical experts reviewed the algorithm in the context of each measure-specific patient cohort and, where clinically indicated, adapted the content of the algorithm to better reflect the likely clinical experience of each measure's patient cohort.
For the THA/TKA readmission measure, CMS used the Planned Readmission Algorithm without making any changes.

The Planned Readmission Algorithm and associated code tables are attached in data field S.2b (Data Dictionary or Code Table). For more details on the Planned Readmission Algorithm, please see the report titled "2016 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures, Version 5.0" posted in data field A.1 or at
https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890567754&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DProcSpecific_Rdmsn_Rpt_2016.pdf&blobcol=urldata&blobtable=MungoBlobs.

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
The target population for the publicly reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures.

Additional details are provided in S.9 Denominator Details.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk, Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
To be included in the measure cohort used in public reporting, patients must meet the following additional inclusion criteria:

1. Enrolled in Medicare fee-for-service (FFS) Part A and Part B Medicare for the 12 months prior to the date of admission; and enrolled in Part A during the index admission;
2. Aged 65 or over;
3. Discharged alive from a non-federal acute care hospital; and,
4. Have a qualifying elective primary THA/TKA procedure; elective primary THA/TKA procedures defined as those procedures without any of the following:

- Femur, hip, or pelvic fractures coded in principal or secondary discharge diagnosis fields of the index admission;
- Partial hip arthroplasty (PHA) procedures with a concurrent THA/TKA;
- Revision procedures with a concurrent THA/TKA;
- Resurfacing procedures with a concurrent THA/TKA;
- Mechanical complication coded in the principal discharge diagnosis field;
- Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field;
- Removal of implanted devices/prostheses; or
- Transfer from another acute care facility for the THA/TKA

This measure can also be used for an all-payer population aged 18 years and older. We have explicitly tested the measure in both patients aged 18 years and older and those aged 65 years or older (see Testing Attachment for details, 2b4.11).

International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes used to define the cohort for each measure are:
ICD-9 codes used to define a THA or TKA:
81.51 Total Hip Arthroplasty
81.54 Total Knee Arthroplasty

ICD-10 codes that define a THA or TKA:
0SR90J9 Replacement of Right Hip Joint with Synthetic Substitute, Cemented, Open Approach
0SR90JA Replacement of Right Hip Joint with Synthetic Substitute, Uncemented, Open Approach
0SR90JZ Replacement of Right Hip Joint with Synthetic Substitute, Open Approach
0SRB0J9 Replacement of Left Hip Joint with Synthetic Substitute, Cemented, Open Approach
0SRB0JA Replacement of Left Hip Joint with Synthetic Substitute, Uncemented, Open Approach
0SRB0JZ Replacement of Left Hip Joint with Synthetic Substitute, Open Approach
0SRC07Z Replacement of Right Knee Joint with Autologous Tissue Substitute, Open Approach
0SRC0JZ Replacement of Right Knee Joint with Synthetic Substitute, Open Approach
0SRC0KZ Replacement of Right Knee Joint with Nonautologous Tissue Substitute, Open Approach
0SRD07Z Replacement of Left Knee Joint with Autologous Tissue Substitute, Open Approach
0SRD0JZ Replacement of Left Knee Joint with Synthetic Substitute, Open Approach

0SRD0KZ Replacement of Left Knee Joint with Nonautologous Tissue Substitute, Open Approach
0SRT07Z Replacement of Right Knee Joint, Femoral Surface with Autologous Tissue Substitute, Open Approach
0SRT0JZ Replacement of Right Knee Joint, Femoral Surface with Synthetic Substitute, Open Approach
0SRT0KZ Replacement of Right Knee Joint, Femoral Surface with Nonautologous Tissue Substitute, Open Approach
0SRU07Z Replacement of Left Knee Joint, Femoral Surface with Autologous Tissue Substitute, Open Approach
0SRU0JZ Replacement of Left Knee Joint, Femoral Surface with Synthetic Substitute, Open Approach
0SRU0KZ Replacement of Left Knee Joint, Femoral Surface with Nonautologous Tissue Substitute, Open Approach
0SRV07Z Replacement of Right Knee Joint, Tibial Surface with Autologous Tissue Substitute, Open Approach
0SRV0JZ Replacement of Right Knee Joint, Tibial Surface with Synthetic Substitute, Open Approach
0SRV0KZ Replacement of Right Knee Joint, Tibial Surface with Nonautologous Tissue Substitute, Open Approach
0SRW07Z Replacement of Left Knee Joint, Tibial Surface with Autologous Tissue Substitute, Open Approach
0SRW0JZ Replacement of Left Knee Joint, Tibial Surface with Synthetic Substitute, Open Approach
0SRW0KZ Replacement of Left Knee Joint, Tibial Surface with Nonautologous Tissue Substitute, Open Approach

An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).

Elective primary THA/TKA procedures are defined as those procedures without any of the following (For a full list of ICD-9 and ICD-10 codes defining the following see attached Data Dictionary, sheet "THA TKA Cohort Codes Part 2"):
1) Femur, hip, or pelvic fractures coded in principal or secondary discharge diagnosis fields of the index admission;
2) Partial hip arthroplasty (PHA) procedures with a concurrent THA/TKA;
3) Revision procedures with a concurrent THA/TKA;
4) Resurfacing procedures with a concurrent THA/TKA;
5) Mechanical complication coded in the principal discharge diagnosis field;
6) Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field;
7) Removal of implanted devises/prostheses; and
8) Transfer status from another acute care facility for the THA/TKA.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
This measure excludes admissions for patients:

1) Without at least 30 days post-discharge enrollment in FFS Medicare;
2) Who were discharged against medical advice (AMA);
3) Admitted for the index procedure and subsequently transferred to another acute care facility;
4) Who had more than two THA/TKA procedure codes during the index hospitalization; or
5) Who had THA/TKA admissions within 30 days of a prior THA/TKA index admission.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
This measure excludes index admissions for patients:

1. Without at least 30 days of post-discharge enrollment in FFS Medicare as determined by examining the Medicare Enrollment Database (EDB).
Rationale: The 30-day readmission outcome cannot be assessed in this group since claims data are used to determine whether a patient was readmitted.

2. Who were discharged against medical advice (AMA), which is identified by examining the discharge destination indicator in claims data.
Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge.

3. Admitted for the index procedure and subsequently transferred to antoher acute care facility, which are defined as when a

patient with an inpatient hospital admission (with at least one qualifying THA/TKA procedure) is discharged from an acute care hospital and admitted to another acute care hospital on the same or next day.
Rationale: Patients admitted for the index procedure and subsequently transferred to another acute care facility are excluded, as determining which hospital the readmission outcome should be attributed to is difficult.

4. Who had more than two THA/TKA procedure codes during the index hospitalization, which is identified by examining procedure codes in the claims data.
Rationale: Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

5. Who had THA/TKA admissions within 30 days prior to THA/TKA index admission.
Rationale: Additional THA/TKA admissions within 30 days are excluded as index admissions because they are part of the outcome. A single admission does not count as both an index admission and a readmission for another index admission.

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
Our approach to risk adjustment is tailored to and appropriate for a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al., 2006).

The measure employs a hierarchical logistic regression model to create a hospital-level 30-day RSRR. In brief, the approach simultaneously models data at the patient and hospital levels to account for the variance in patient outcomes within and between hospitals (Normand & Shahian, 2007). At the patient level, the model adjusts the log-odds of readmission within 30 days of discharge for age and selected clinical covariates. At the hospital level, the approach models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of readmission at the hospital, after accounting for patient risk. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Candidate and Final Risk-adjustment Variables: Candidate variables were patient-level risk-adjustors that were expected to be predictive of readmission, based on empirical analysis, prior literature, and clinical judgment, including age and indicators of comorbidity and disease severity. For each patient, covariates are obtained from claims records extending 12 months prior to and including the index admission. For the measure currently implemented by CMS, these risk adjusters are identified using both inpatient and outpatient Medicare FFS claims data. However, in the all-payer hospital discharge database measure, the risk-adjustment variables can be obtained only from inpatient claims in the prior 12 months and the index admission.

The model adjusts for case-mix differences based on the clinical status of patients at the time of admission. We use condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes (Pope et al., 2000). A file that contains a list of the ICD-9-CM codes and their groupings into CCs is attached in data field S.2b (Data Dictionary or Code Table). In addition, only comorbidities that convey information about the patient at admission or in the 12 months prior, and not complications that arise during the course of the index hospitalization, are included in the risk adjustment. Hence, we do not risk adjust for CCs that may represent adverse events of care when they are only recorded in the index admission.

The final set of risk-adjustment variables is:

Demographics
Age-65 (years, continuous) for patients aged 65 or over cohorts; or Age (years, continuous) for patients aged 18 and over cohorts
Male (%)

THA/TKA Procedure
Index admissions with an elective THA procedure
Number of procedures (two vs. one)

Clinical Risk Factors
Other congenital deformity of hip (joint) (ICD-9 code 755.63)
Post traumatic osteoarthritis (ICD-9 codes 716.15, 716.16)
Morbid obesity (ICD-9 code 278.01)
History of infection (CC 1, 3-6)
Metastatic cancer or acute leukemia (CC 7)
Cancer (CC 8-12)
Diabetes mellitus (DM) or DM complications (CC 15-20, 119-120)
Protein-calorie malnutrition (CC 21)
Disorders of fluid/electrolyte/acid-base (CC 22-23)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)
Severe hematological disorders (CC 44)
Dementia or other specified brain disorders (CC 49, 50)
Major psychiatric disorders (CC 54-56)
Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)
Polyneuropathy (CC 71)
Congestive heart failure (CC 80)
Coronary atherosclerosis or angina (CC 83-84)
Hypertension (CC 89, 91)
Specified arrhythmias and other heart rhythm disorders (CC 92-93)
Stroke (CC 95-96)
Vascular or circulatory disease (CC 104-106)
Chronic obstructive pulmonary disease (COPD) (CC 108)
Pneumonia (CC 111-113)
Dialysis status (CC 130)
Renal failure (CC 131)
Decubitus ulcer or chronic skin ulcer (CC 148-149)
Cellulitis, local skin infection (CC 152)
Other injures (CC 162)
Major symptoms, abnormalities (CC 166)

References:
Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*

Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Available in attached Excel or csv file at S.2b

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
N/A

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
The measure estimates hospital-level 30-day all-cause RSRRs following elective primary THA/TKA using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals (Normand and Shahian, 2007). At the patient level, it models the log-odds of readmission within 30 days of discharge using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, it models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a readmission at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

The RSRR is calculated as the ratio of the number of "predicted" to the number of "expected" readmission at a given hospital, multiplied by the national observed readmission rate. For each hospital, the numerator of the ratio is the number of readmissions within 30 days predicted on the basis of the hospital's performance with its observed case mix, and the denominator is the number of readmissions expected based on the nation's performance with that hospital's case mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows for a comparison of a particular hospital's performance given its case mix to an average hospital's performance with the same case mix. Thus, a lower ratio indicates lower-than-expected readmission rates or better quality, and a higher ratio indicates higher-than-expected readmission rates or worse quality.

The "predicted" number of readmissions (the numerator) is calculated by using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of readmission. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of readmissions (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the years of data in that period.

This calculation transforms the ratio of predicted over expected into a rate that is compared to the national observed readmission rate. The hierarchical logistic regression models are described fully in the original methodology report (Grosso et al., 2012).

References:
Grosso L, Curtis J, Geary L, et al. Hospital-level 30-Day All-Cause Risk-Standardized Readmission Rate Following Elective Primary Total Hip Arthroplasty (THA) And/Or Total Knee Arthroplasty (TKA) Measure Methodology Report. 2012.

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22(2): 206-226.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A. This measure is not based on a sample.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A. This measure is not based on a survey or patient-reported data.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing values are rare among variables used from claims data in this measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Administrative claims, Other

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Data sources:

The currently publically reported measure is specified and has been testing using:
1. Medicare Part A inpatient and Part B outpatient claims: This data source contains claims data for FFS inpatient and outpatient services including: Medicare inpatient hospital care, outpatient hospital services, as well as inpatient and outpatient physician claims for the 12 months prior to an index admission.

2. Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission as well as vital status at discharge. These data have previously been shown to accurately reflect patient vital status (Fleming et al., 1992).

The measure was also specified and testing using an all-payer claims dataset although it is only publically reported using the data sources listed above:
3. California Patient Discharge Data in addition to CMS Medicare FFS data for patients in California hospitals. Using all-payer data from California, we performed analyses to determine whether the THA/TKA readmission measure can be applied to all adult patients, including not only FFS Medicare patients aged 65 years or over, but also non-FFS Medicare patients aged 18-64 years at the time of admission.

Additional data source used for the analysis of the impact of SES variables on the measure's risk model. Note that the variables derived from these data are not included in the measure as specified
4. The American Community Survey (2009-2013): The American Community Survey data is collected annually and an aggregated 5-years data was used to calculate the AHRQ socioeconomic status (SES) composite index score.

Reference:
Fleming C., Fisher ES, Chang CH, Bubolz D, Malenda J. Studying outcomes and hospital utilization in the elderly: The advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Medical Care. 1992; 30(5): 377-91.

Dorsey K, Grady J, Desai N, et al. 2016 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures: Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) & Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Version 5.0). 2016

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
N/A. This measure is not a composite performance measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
NQF_1551_HipKnee_Readmission_NQF_Testing_Attachment_v1.0.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 1551
Measure Title: Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
Date of Submission: 5/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*)
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.

- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions
15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: **(*must be consistent with data sources entered in S.23*)**	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☒ other: Census Data/American Community Survey

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

The datasets used for testing included Medicare Parts A and B claims, as well as the Medicare Enrollment Database (EDB). Additionally, census data were used to assess socioeconomic factors and race (dual eligibility and African American race variables obtained through enrollment data; Agency for Healthcare Research and Quality [AHRQ] socioeconomic status [SES] index score obtained through census data). Data abstracted from hospital medical records were used to validate the claims-based assessment of the readmission outcome. The dataset used varies by testing type; see Section 1.7 for details.

1.3. What are the dates of the data used in testing?

The dates used vary by testing type; see Section 1.7 for details.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: **(*must be consistent with levels entered in item***	**Measure Tested at Level of:**

S.26)	
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

For this measure, hospitals are the measured entities. All non-federal, acute inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The number of admissions/patients varies by testing type: see Section 1.7 for details

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The datasets, dates, number of measured entities and number of admissions used in each type of testing are as follows:

For reliability testing (Section 2a2)

For reliability testing, we randomly split **Dataset 1** into two samples. The reliability of the model was tested by randomly selecting 50% of the Medicare patients aged 65 years and over in the most recent three-year cohort and developing a risk-adjusted model for this group. We then developed a second model for the remaining 50% of patients and compared the two. In each year of measure reevaluation, we also re-fit the model and compared the frequencies and model coefficients of risk variables (condition categories for patient comorbidities) and model fit across 3 years (**Dataset 1** below).

Dataset 1 (2015 public reporting cohort): Medicare Part A Inpatient and Outpatient and Part B Outpatient claims, and Medicare Enrollment Database (to assess enrollment in FFS)
Dates of Data: July 1, 2011 – June 30, 2014
Number of Index Admissions: 926,685
Patient Descriptive Characteristics: average age= 74.7, %male= 36.8

Number of Measured Entities: 3,498 hospitals

For validity testing (Section 2b2)

No empirical testing was done

For testing of measure exclusions (Section 2b3)

Dataset 1 (2015 public reporting cohort)

For testing of measure risk adjustment (Section 2b4)

Dataset 1 (2015 public reporting cohort)

Dataset 2 (development dataset): Medicare Part A Inpatient and Outpatient, and Part B Outpatient claims
Dates of Data: Calendar year (January – December) 2008
Number of Admissions: N=148,132 (first half of split sample); N=148,092 (second half of split sample)

Number of Measured Entities: 3,223 hospitals (first half of split sample); 3,213 hospitals (second half of split samples)

To create the model development sample (**Dataset 2**), we applied the inclusion and exclusion criteria to all 2008 admissions. We randomly selected half of all THA/TKA admissions in 2008 that met the inclusion and exclusion criteria to create a model development sample. We used the remaining admissions as our model validation sample.

For Sub-section 2b4.11. Optional Additional Testing for Risk Adjustment

We performed additional testing of the measure in an all-payer claims dataset

Dataset 3 (all payer dataset, section 2b4.11): California Patient Discharge Data in addition to CMS Medicare FFS data for patients in California hospitals
Dates of Data: January 1, 2006 – December 31, 2006
Number of Index Admissions: 61,831 (all patients 18 years and older) [mean age=67.1, %male=39.8]

For testing to identify meaningful differences in performance (Section 2b5)

Dataset 1 (2015 public reporting cohort)

For testing of sociodemographic factors in risk models (Section 2b4.4b)

Dataset 1 (2015 public reporting cohort); **Dataset 4** (The American Community Survey [ACS]): The American Community Survey, 2009-2013

We examined disparities in performance according to the proportion of patients in each hospital who were of African-American race and the proportion who were dual eligible for both Medicare and Medicaid insurances. We also used the AHRQ SES index score to study the association between performance measures and socioeconomic status.

Data Elements

• African-American race and dual eligible status (i.e., enrolled in both Medicare and Medicaid) patient-level data are obtained from CMS enrollment data (**Dataset 1**)

• Validated AHRQ SES index score is a composite of 7 different variables found in the census data (**Dataset 4**)

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

SDS incorporates socioeconomic variables as well as race into a more concise term. However, given the fact that socioeconomic risk factors are distinct from race and should be interpreted differently, we have decided to keep "socioeconomic status" and "race" as separate terms.

We selected socioeconomic status (SES) and race variables to analyze after reviewing the literature and examining available national data sources. There is a large body of literature linking various SES factors and African-American race to worse health status and higher readmission risk (Blum et al., 2014; Eapen et al. 2015; Gilman et al., 2014; Hu et al., 2014; Joynt and Jha, 2013). Income, education, and occupational level are the most commonly examined variables. Although literature directly examining how different SES factors or race might influence the likelihood of older, insured, Medicare patients being readmitted within 30 days of an admission for THA/TKA is much more limited, available studies have indicated an increased risk of readmission (Oronce et al., 2015; Singh et al., 2014). However, others have found SES and race may not predict 30-day readmissions after orthopedic procedures (Hunter et al., 2015). In addition, studies have also suggested other disparities related to hip/knee surgery, including significant differences in the rate of total hip replacement surgery received by African-American and white patients (Ibrahim, 2010; Mahomed, 2003). The causal pathways for SES and race variable selection are described below in Section 2b4.3.

The SES and race variables used for analysis were:

- Dual eligible status (**Dataset 1**)
- African-American race (**Dataset 1**)
- AHRQ-validated SES index score (summarizing the information from the following variables: percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th-grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room) (**Dataset 4**)

In selecting variables, our intent was to be responsive to the NQF guidelines for measure developers in the context of the SDS Trial Period. Our approach has been to examine all patient-level indicators of both SES and race that are reliably available for all Medicare beneficiaries, are linkable to claims data, and have established validity.

Previous studies examining the validity of data on patients' race and ethnicity collected by CMS have shown that only the data identifying African-American beneficiaries have adequate sensitivity and specificity to be applied broadly in research or measures of quality. While using this variable is not ideal because it groups all non-African-American beneficiaries together, it is currently the only race variable available on all beneficiaries across the nation that is linkable to claims data.

We similarly recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous outcome. However, the threshold for over 65-year-old Medicare patients is valuable, as it takes into account both income and assets and is consistently applied across states.

For both our race and the dual-eligible variables, there is a body of literature demonstrating differential health care and health outcomes among beneficiaries indicating that these variables, while not ideal, also allow us to examine some of the pathways of interest.

Finally, we selected the AHRQ-validated SES index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas (Bonito et al., 2008). Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. In this submission, we present analyses using the census block level, the most granular level possible using ACS data. A census block group is a geographical unit used by the US Census Bureau which is between the census tract and the census block. It is the smallest geographical unit for which the bureau publishes sample data. The target size for block groups is 1,500 people and they typically have a population of 600 to 3,000 people. We used 2009-2013 ACS data and mapped patients' 9-digit ZIP codes via vendor software to the AHRQ SES Index at the census block group level. Given the variation in cost of living across the country, the median income and median property value components of the AHRQ SES Index were adjusted by regional price parity values published by the Bureau of Economic Analysis (BEA). This provides a better marker of low SES neighborhoods in high expense geographic areas. We then calculated an AHRQ SES Index score for census block groups that can be linked to 9-digit ZIP codes. In the THA/TKA measure cohort, we were able to assign an AHRQ SES Index score to 99.6% of patient admissions. 88.6% of patient admissions had calculated AHRQ SES Index scores linked to their 9-digit ZIP codes. 11.0% of patient admissions had only valid 5-digit ZIP codes; we utilized the data for the median 9-digit ZIP code within that 5-digit ZIP code.

References:

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation. Cardiovascular quality and outcomes. May 2014; 7(3):391-397.

Bonito A, Bann C, Eicheldinger C, Carpenter L. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final Report, Sub-Task. 2008;2.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, HernandezAF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

Hunter T, Yoon RS, Hutzler L, et al. No evidence for race and socioeconomic status as independent predictors of 30-day readmission rates following orthopedic surgery. American Journal of Medical Quality: The Official Journal of the American College of Medical Quality. 2015;30(5):484-488.

Ibrahim SA. Racial variations in the utilization of knee and hip joint replacement: an introduction and review of the most recent literature. Current orthopaedic practice 2010;21:126-131

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. The Journal of bone and joint surgery American volume 2003;85-a:27-32

Oronce CI, Shao H, Shi L. Disparities in 30-Day Readmissions After Total Hip Arthroplasty. Medical care. 2015;53(11):924-930.

Singh JA, Lu X, Rosenthal GE, Ibrahim S, Cram P. Racial disparities in knee and hip total joint arthroplasty: an 18-year analysis of national Medicare data. Annals of the rheumatic diseases. 2014;73(12):2107-2115.

2a2. RELIABILITY TESTING

Note: *If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Data Element Reliability

In constructing the measure, we aim to utilize only those data elements from the claims that have both face validity and reliability. We avoid the use of fields that are thought to be coded inconsistently across hospitals or providers. Specifically, we use fields that are consequential for payment and which are audited. We identify such variables through empiric analyses and our understanding of CMS auditing and billing policies and seek to avoid variables which do not meet this standard. For example, "discharge disposition" is a variable in Medicare claims data that is not thought to be a reliable variable for identifying a transfer between two acute care facilities. Thus, we derive a variable using admission and discharge dates as a surrogate for "discharge disposition" to identify hospital admissions involving transfers. This allows us to identify these admissions using variables in the claims data that have greater

reliability than the “discharge disposition” variable.

In addition, CMS has in place several hospital auditing programs used to assess overall claims code accuracy, to ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes and other elements that are consequential to payment.

Finally, we assess the reliability of the data elements by comparing model variable frequencies and odds ratios from logistic regression models across the most recent three years of data (**Dataset 1**, see section 1.7).

Measure Score Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. In line with this thinking, our approach to assessing reliability was to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produces similar measures of hospital performance. That is, we took a "test-retest" approach in which hospital performance was measured once using a random subset of patients, then measured again using a second random subset exclusive of the first. Finally, we compared the agreement between the two resulting performance measures across hospitals (Rousson et al., 2002).

For test-retest reliability, we combined index admissions from successive measurement periods into one dataset, randomly sampled half of patients within each hospital, calculated the measure for each hospital, and repeated the calculation using the second half of patients. Thus, each hospital was measured twice, but each measurement was made using an entirely distinct set of patients. To the extent that the calculated measures of these two samples agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement, we calculated the intra-class correlation coefficient (ICC) (Shrout and Fleiss, 1979), and assessed the values according to conventional standards (Landis and Koch, 1977). Specifically, we used the **Dataset 1** split sample and calculated the RSRR for each hospital for each sample. The agreement of the two RSRRs was quantified for hospitals using the ICC (2,1) as defined by Shrout and Fleiss (1979).

Using two independent samples provides a stringent estimate of the measure’s reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less ´signal´, a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910). We used this formula to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

References:

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Landis J, Koch G. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. Statistics in Medicine 2002;21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 1979;86:420-428.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Data Element Reliability Results (**Dataset 1**)

The frequency of some model variables increased while others decreased between 2011 and 2014, which may reflect an increased or decreased rate of specific comorbidities in the FFS population. For example, there was a notable decrease in the frequency (≥ 2%) for index admissions with Coronary Atherosclerosis or Angina (CC 83-84) (from 28.7% to 26.4%). Examination of the odds ratios for each risk variable in the model shows that, overall, the odds ratios for individual risk variables remained relatively constant across the three years. See the *2015 Measure Updates and Specifications Report* for details (Dorsey et al., 2015).

Measure Score Reliability Results (**Dataset 1**)

There were 926,685 admissions in the 3-year split sample (from **Dataset 1**), with 460,576 index admissions from 2,835 hospitals in one sample and 459,237 admissions from 2,835 hospitals in the other randomly selected sample. The agreement between the two RSRRs for each hospital, the ICC, was 0.49, which according to the conventional interpretation is "moderate" (Landis & Koch, 1977).

Note that we limited this analysis to hospitals with 12 or more cases in each split sample.

The ICC is based on a split sample of three years of data, resulting in a volume of patients in each sample equivalent to only 1.5 years of data, whereas the measure is reported with the full three years of data.

References:

Dorsey K, Grady J, Desai N, et al. 2016 Procedure-Specific Measures Updates and Specifications Report Hospital-Level 30-Day Risk-Standardized Readmission Measures: Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) & Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Version 5.0). 2016 https://www.qualitynet.org/dcs/BlobServer?blobkey=id&blobnocache=true&blobwhere=1228890567754&blobheader=multipart%2Foctet-stream&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3Bfilename%3DProcSpecific_Rdmsn_Rpt_2016.pdf&blobcol=urldata&blobtable=MungoBlobs. Accessed May 16, 2016.

Landis J, Koch G. The measurement of observer agreement for categorical data, Biometrics 1977;33:159-174.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The stability of the risk factor odds ratios over time suggests that the underlying data elements are reliable. Additionally, the ICC score demonstrates moderate agreement across samples using a conservative approach to assessment.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☐ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Measure validity is demonstrated through prior validity testing done on our claims-based measures, through use of established measure development guidelines, and by systematic assessment of measure face validity by a technical expert panel (TEP) of national experts and stakeholder organizations.

Validity of Claims-Based Measures:

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated six NQF-endorsed measures currently in public reporting (heart failure, acute myocardial infraction [AMI], and pneumonia mortality and readmission measures) with models that used chart-abstracted data for risk adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical record data for risk adjustment for heart failure patients (National Heart Failure data) (Krumholz et al. 2006; Keenan et al. 2008), AMI patients (Cooperative Cardiovascular Project data) (Krumholz, Wang, et al. 2006), and pneumonia patients (National Pneumonia Project dataset) (Bratzler et al. 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (elective primary THA/TKA and implantable cardioverter defibrillator). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical record data.

Validity Indicated by Established Measure Development Guidelines:

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz, Brindis, et al. 2006).

Validity as Assessed by External Groups:

Throughout measure development, we obtained expert and stakeholder input via three mechanisms: regular discussions with an advisory working group, a national TEP, and a 15-day public comment period in order to increase transparency and to gain broader input into the measure.

We assembled the working group and held regular meetings throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians and other professionals with expertise in biostatistics, measure methodology, and quality improvement. Working group meetings addressed key issues related to measure development, including weighing the pros and cons of and finalizing key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader TEP.

In addition to the working group, and in alignment with the CMS MMS, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including physicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS site link. The public comments received during the call for comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

Finally, NQF previously endorsed this measure in 2012, demonstrating additional external groups' endorsement of the measure's validity.

Face Validity as Determined by TEP:

One means of confirming the validity of this measure was face validity assessed by our TEP.

List of TEP Members:

Mark L. Francis, MD
Professor of Medicine and Biomedical Sciences, Chief, Division of Rheumatology, Department of Internal Medicine, Texas Tech University Health Sciences Center

Cynthia Jacelon, PhD, RN, CRRN
Associate Professor, School of Nursing, University of Massachusetts
Association of Rehabilitation Nurses

Norman Johanson, MD
Chairman, Orthopedic Surgery, Drexel University College of Medicine

C. Kent Kwoh, MD
Professor of Medicine, Associate Chief and Director of Clinical Research, Division of Rheumatology and Clinical Immunology University of Pittsburgh

Courtland G. Lewis, MD
American Association of Orthopaedic Surgeons

Jay Lieberman, MD
Professor and Chairman, Department of Orthopedic Surgery, University of Connecticut Health Center; Director, New England Musculoskeletal Institute

Peter Lindenauer, MD, M.Sc.
Hospitalist and Health Services Researcher, Baystate Medical Center; Professor of Medicine, Tufts University

Russell Robbins, MD, MBA
Principal, Mercer's Total Health Management

Barbara Schaffer
THA Patient

Nelson SooHoo, MD, MPH
Professor, University of California at Los Angeles

Steven H. Stern, MD
Vice President, Cardiology & Orthopedics/ Neuroscience, UnitedHealthcare

Richard E. White, Jr., MD
American Association of Hip and Knee Surgeons

Citations:

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Brindis RG,Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.

Krumholz HM, Brindis RG,Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

ICD-9 to ICD-10 Conversion

Statement of Intent

[X] Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.
[ ] Goal was to take advantage of the more specific code set to form a new version of the measure, but fully consistent with the original intent.
[ ] The intent of the measure has changed.

Process of Conversion

ICD-10 codes were initially identified using General Equivalence Mapping (GEM) software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes currently in use for this measure. Each year we reexamine the codes using the latest version of the GEM software. We completed this examination most recently in 2015. An ICD-9 to ICD-10 crosswalk is attached in field S.2b. (Data Dictionary or Code Table).

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Validity as Assessed by External Groups:

The TEP, comprised of individuals with expertise relevant to orthopedic quality measurement, provided input on the model to strengthen the measure and supported the final measure.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Validity as Assessed by External Groups:

The TEP's feedback on the measure demonstrated their agreement with the overall face validity of the measure as specified.

2b3. EXCLUSIONS ANALYSIS
NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions and to ensure accurate calculation of the measure. To ascertain the impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Dataset 1**). These exclusions are consistent with similar NQF-endorsed outcome measures. For more details, see the attached specifications report.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

In **Dataset 1** (2015 public reporting cohort):

Exclusion	N	%	Distribution across hospitals (N=2,826): Minimum, 25th percentile, 50th percentile, 75th percentile, maximum
1. Discharged against medical advice (AMA)	112	0.01	(0.0, 0.0, 0.0, 0.0, 0.3)
2. Without at least 30 days post-discharge enrollment in FFS Medicare for index admissions	1,892	0.20	(0.0, 0.0, 0.0, 0.2, 5.5)
3. Admitted for the index procedure and subsequently transferred to another acute care facility	8,515	0.91	(0.0, 0.0, 0.0, 0.5, 7.5)
4. Who had more than two THA/TKA procedure codes during the index hospitalization	1	0.00	(0.0, 0.0, 0.0, 0.0, 0.2)
5. Who had an admission for THA/TKA within 30 days of a prior index admission	1,397	0.15	(0.0, 0.0, 0.0, 0.0, 8.6)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Exclusion 1 (patients who are discharged AMA) accounts for 0.01% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the opportunity to adequately deliver full care and prepare the patient for discharge. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 2 (patients without at least 30 days of post-discharge enrollment in FFS Medicare for index admissions) accounts for 0.20% of all index admissions excluded from the initial cohort. This exclusion is needed because the 30-day readmission outcome cannot be assessed in this

group since claims data are used to determine whether a patient was readmitted. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 3 (patients who are transferred to another acute care facility) accounts for 0.91% of all index procedures excluded from the initial index cohort. This exclusion is intended to remove admissions from the cohort for patients transferred in to the index hospital, as they likely do not represent elective THA/TKA procedures.

Exclusion 4 (patients with more than two THA/TKA procedure codes during the index hospitalization) accounts for 0.00% of all index admissions excluded from the initial index cohort. This exclusion is needed to ensure a clinically coherent cohort. Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one admission, which may reflect a coding error.

Exclusion 5 (patients who had an admission within 30 days of a prior index admission) accounts for 0.15% of all index admissions excluded from the initial index cohort. This exclusion is needed to prevent admissions from being counted as both an index admission and a readmission.

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

I***f not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.***

2b4.1. What method of controlling for differences in case mix is used?

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 33 risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Our approach to risk adjustment was tailored to, and appropriate for, a publicly reported outcome measure, as articulated in the American Heart Association (AHA) Scientific Statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz et al. 2006).

The measure employs a hierarchical logistic regression model (a form of hierarchical generalized linear model [HGLM]) to create a hospital-level 30-day RSRR. This approach to modeling appropriately accounts for the structure of the data (patients clustered within hospitals), the underlying risk due to patients' comorbidities, and sample size at a given hospital when estimating hospital readmission rates. In brief, the approach simultaneously models two levels (patient and hospital) to account for the variance in patient outcomes within and between hospitals (Normand and Shahian et al. 2007). At the patient level, each model adjusts the log odds of readmission within 30-days of admission for age, sex, selected clinical covariates and a hospital-specific intercept. The second level models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept, or hospital-specific effect, represents the hospital contribution to the risk of readmission, after accounting for patient risk and sample size, and can be inferred as a measure of quality. The hospital-specific intercepts are given a distribution in order to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

Clinical Factors

Candidate and Final Risk-adjustment Variables

The original measure was developed using Medicare FFS claims data. Candidate variables were patient-level risk adjustors that were expected to be predictive of readmission, based on empirical analysis, prior literature, and clinical judgment, including demographic factors (age, sex) and indicators of comorbidity and disease severity. For each patient, covariates were obtained from Medicare claims extending 12 months prior to and including the index admission. The model adjusted for case differences based on the clinical status of the patient at the time of admission. We used condition categories (CCs), which are clinically meaningful groupings of more than 15,000 ICD-9-CM diagnosis codes. We did not risk adjust for CCs that were possible adverse events of care and that were only recorded in the index admission. In addition, only comorbidities that conveyed information about the patient at that time or in the 12 months prior, and not complications that arose during the course of the admission were included in the risk adjustment.

The final set of risk-adjustment variables is:

Demographic

- Age-65 (years above 65, continuous)
- Male

THA/TKA Procedure

- Index admissions with an elective THA procedure
- Number of procedures (two vs. one)

Clinical Risk Factors

- Other congenital deformity of hip (joint) ICD-9 code 755.63
- Post traumatic osteoarthritis ICD-9 codes 716.15, 716.16
- Morbid obesity ICD-9 code 278.01
- History of infection (CC 1, 3-6)
- Metastatic cancer and acute leukemia (CC 7)

- Cancer (CC 8-12)
- Diabetes mellitus (DM) or DM complications (CC 15-20, 119-120)
- Protein-calorie malnutrition (CC 21)
- Disorders of fluid/electrolyte/acid-base (CC 22-23)
- Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)
- Severe hematological disorders (CC 44)
- Dementia or other specified brain disorder (CC 49-50)
- Major psychiatric disorders (CC 54-56)
- Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)
- Polyneuropathy (CC 71)
- Congestive heart failure (CC 80)
- Coronary atherosclerosis or angina (CC 83-84)
- Hypertension (CC 89, 91)
- Specified heart arrhythmias and other heart rhythm disorders (CC 92-93)
- Stroke (CC 95-96)
- Vascular or circulatory disease (CC 104-106)
- Chronic obstructive pulmonary disease (COPD) (CC 108)
- Pneumonia (CC 111-113)
- Dialysis status (CC 130)
- Renal failure (CC 131)
- Decubitus ulcer or chronic skin ulcer (CC 148-149)
- Cellulitis, local skin infection (CC 152)
- Other injuries (CC 162)
- Major symptoms, abnormalities (CC 166)

Socioeconomic Status (SES) Factors and Race

We selected variables representing SES factors and race for examination based on a review of literature, conceptual pathways, and feasibility. In Section 1.8, we describe the variables that we considered and analyzed based on this review. Below we describe the pathways by which SES and race may influence 30-day readmission.

Our conceptualization of the pathways by which patient SES or race affects 30-day readmission is informed by the literature.

Literature Review of Socioeconomic Status (SES) and Race Variables and THA/TKA Readmission

To examine the relationship between SES and race variables and hospital 30-day RSRR following elective primary THA/TKA, a literature search was performed with the following exclusion criteria: international studies, articles published more than 10 years ago, articles without primary data, articles using Veterans Affairs databases as the primary data source, and articles not explicitly focused on SES or race and hip/knee surgery readmission. Six studies were initially reviewed, and three studies were excluded from full-text review based on the above criteria. The studies available for review indicated that SES and race variables may be associated with increased risk of THA/TKA readmission (Oronce et al., 2015; Singh et al., 2014), though the evidence was very limited with one study suggesting that neither race nor SES predicted risk of readmission (Hunter et al., 2015). In addition to the literature focused on readmissions following hip/knee surgery, other studies have also found significant differences in the rate of

THA received by African-American and white patients, indicating that patient and surgeon behavior may also contribute to disparities based on racial factors (Ibrahim, 2010; Mahomed et al., 2003).

Causal Pathways for Socioeconomic Status (SES) and Race Variable Selection

Although some recent literature evaluates the relationship between patient SES or race and the readmission outcome across conditions and procedures, few studies directly address causal pathways or examine the role of the hospital in these pathways. Moreover, the current literature examines a wide range of conditions and risk variables with no clear consensus on which risk factors demonstrate the strongest relationship with readmission. The SES factors that have been examined in the readmission literature, which spans across conditions and procedures, can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables. Patient-level variables describe characteristics of individual patients, and range from the self-reported or documented race or ethnicity of the patient to the patient's income or education level (Eapen et al., 2015; Hu et al., 2014; Ibrahim, 2010). Neighborhood/community-level variables use information from sources such as the American Community Survey (ACS) as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the Agency for Healthcare Research and Quality (AHRQ)-validated SES index score (Blum et al., 2014). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman et al., 2014; Joynt and Jha, 2013).

The conceptual relationship, or potential causal pathways by which these possible SES risk factors influence the risk of readmission following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider.

1. **Relationship of socioeconomic status (SES) factors or race to health at admission**. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their index admission or procedure with a greater severity of underlying illness. These SES risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job, lack of childcare), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all may lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk adjustment**.**

In addition to SES risk factors, studies have shown that worse health status is more prevalent among African-American patients compared with white patients. The association between race and worse health is in part mediated by the association between race and SES risk factors such as poverty or disparate access to care associated with poverty or neighborhood. The association is also mediated through bias in health care as well as other facets of society.

2. **Use of low-quality hospitals**. Patients of lower income, lower education, or unstable housing have been shown not to have equitable access to high-quality facilities because such facilities are less likely to be found in geographic areas with large populations of poor patients; thus, patients with low income are more likely to be seen in lower quality hospitals, which can contribute to

increased risk of readmission (Jha et al., 2011; Reames et al., 2014). Similarly African-American patients have been shown to have less access to high-quality facilities compared with white patients (Skinner et al., 2005).

3. **Differential care within a hospital**. The third major pathway by which SES factors or race may contribute to readmission risk is that patients may not receive equivalent care within a facility. For example, African-American patients have been shown to experience differential, lower quality, or discriminatory care within a given facility (Trivedi et al., 2014). Alternatively, patients with SES risk factors such as lower education may require differentiated care – e.g. provision of lower literacy information – that they do not receive.

4. **Influence of SES on readmission risk outside of hospital quality and health status**. Some SES risk factors, such as income or wealth, may affect the likelihood of readmission without directly affecting health status at admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing economic priorities or a lack of access to care outside of the hospital.

These proposed pathways are complex to distinguish analytically. They also have different implications on the decision to risk adjust or not. We, therefore, first assessed if there was evidence of a meaningful effect on the risk model to warrant efforts to distinguish among these pathways. Based on this model and the considerations outlined in Section 1.8, the following SES and race variables were considered:

• Dual-eligible status,
• African-American race, and
• AHRQ SES index

We assessed the relationship between the SES and race variables with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

One concern with including SES or race factors in a model is that their effect may be at either the patient or the hospital level. For example, low SES may increase the risk of readmission because patients of low SES have an individual higher risk (patient-level effect) or because patients of low SES are more often admitted to hospitals with higher overall readmission rates (hospital-level effect). **Identifying the relative contribution of the hospital-level effect is important in considering whether a factor should be included in risk adjustment; if an effect is primarily a hospital-level effect, adjusting for it is equivalent to adjusting for differences in hospital quality.** Thus, as an additional step, we assessed whether there was a "contextual effect" at the hospital level. To do this, we performed a decomposition analysis to assess the independent effects of the SES and race variables at the patient level and the hospital level. If, for example, all the elevated risk of readmission for patients of low SES were due to lower quality/higher readmission risk in hospitals with more patients of low SES, then a significant hospital-level effect would be expected with little-to-no patient-level effect. However, if the increased readmission risk was solely related to higher risk for patients of low SES regardless of hospital effect, then a significant patient-level effect would be expected and a significant hospital-level effect would not be expected.

Specifically, we modeled each of the SES and race variables as follows: Let X_{ij} be a binary indicator of SES or race of the i^{th} patient at the j^{th} hospital, and X_j the percent of patients at hospital j with $X_{ij} = 1$. Then, we added both $X_{ij} \equiv X_{patient}$ and $X_j \equiv X_{hospital}$ to the model. The first variable, $X_{patient}$, represents the effect of the risk factor at the patient level (sometimes called the "within" hospital effect), and the second variable, $X_{hospital}$, represents the effect at the hospital level (sometimes called the "between" hospital effect). By including both of these in the same model, we assessed whether these were independent effects, whether one effect dominated the other, or whether only one of these effects contributed. This analysis allowed us to simultaneously estimate the independent effects of: 1) hospitals with higher or lower proportions of low SES patients or African-American patients on the readmission rate of an average patient; and 2) a patient's SES or race on his or her own readmission rate when seen at an average hospital.

It is very important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect.

It is also important to note that the patient-level and hospital-level coefficients cannot be quantitatively compared because the patient's SES or race in the model is binary, whereas the hospitals' proportion of low SES patients or African-American patients is continuous. Therefore, in order to quantitatively compare the relative size of the patient- and hospital-level effects, we calculated a range of predicted probabilities of readmission based on the fitted model.

Specifically, to estimate an average hospital effect, we calculated the predicted probabilities for the following scenarios: (1) Assuming all patients do not have the risk factor (X_{ij} =0) and hospital-level risk factor is at 5% percentile (P5) of all hospital values; (2) Assuming all patients do not have the risk factor and hospital-level risk factor is at 95% percentile (P95); (3) Assuming all patients have the risk factor (X_{ij} =1) and hospital-level risk factor is at 5% percentile (P5); (4) Assuming all patients have the risk factor and hospital-level risk factor is at 95% percentile (P95). The average hospital-level effect was estimated by ((2)-(1) + (4)-(3))/2 (P95-P5). Then, to estimate an average patient-level effect, we first calculated the predicted probabilities by assuming patient-level risk factor equal to 0 or 1 at different hospital risk factor percentiles (0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%). Then, at each of those percentiles, we obtained the difference of predicted probabilities between all patients not having the risk factor and all patients having the risk factor. We calculated the average of those differences in predicted probabilities ('delta') as the patient-level effect.

In summary, the difference in predicted probabilities at the 95^{th} and 5^{th} percentiles (P95-P5) estimates the hospital-level effect of the SES or race risk factor on readmission. The difference in predicted probabilities when all patients have and do not have the SES or race risk factor (delta) estimates the patient-level effect of the SES or race risk factor on readmission. The hospital-level effect is greater than the patient-level effect when P95-P5 is greater than delta. We used P95 and P5 rather than the maximum (P100) and minimum (P0) to avoid outlier values.

We also performed the same analysis for several clinical covariates to contrast the relative contributions of patient- and hospital-level effects of clinical variables to the relative

contributions for the SES and race variables. Refer to section 2b4.4b for the results of these analyses.

References:

Blum, A. B., N. N. Egorova, E. A. Sosunov, A. C. Gelijns, E. DuPree, A. J. Moskowitz, A. D. Federman, D. D. Ascheim and S. Keyhani. "Impact of Socioeconomic Status Measures on Hospital Profiling in New York City." *Circ Cardiovasc Qual Outcomes* 7, no. 3 (2014): 391-7.

Eapen ZJ, McCoy LA, Fonarow GC, Yancy CW, Miranda ML, Peterson ED, Califf RM, Hernandez AF. Utility of socioeconomic status in predicting 30-day outcomes after heart failure hospitalization. Circ Heart Fail. May 2015; 8(3):473-80.

Gilman M, Adams EK, Hockenberry JM, Wilson IB, Milstein AS, Becker ER. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health affairs (Project Hope). 2014; 33(5):778-785.

Hunter T, Yoon RS, Hutzler L, et al. No evidence for race and socioeconomic status as independent predictors of 30-day readmission rates following orthopedic surgery. American Journal of Medical Quality: The Official Journal of the American College of Medical Quality. 2015;30(5):484-488.

Ibrahim SA. Racial variations in the utilization of knee and hip joint replacement: an introduction and review of the most recent literature. Current orthopaedic practice 2010;21:126-131

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and Medicaid patients. Health Affairs 2011; 30:1904-11.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Krumholz HM, Brindis RG, Brush JE, et al. 2006. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation 113: 456-462.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. The Journal of bone and joint surgery American volume 2003;85-a:27-32

Normand S-LT, Shahian DM. 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

Reames BN, Birkmeyer NJ, Dimick JB, Ghaferi AA. Socioeconomic disparities in mortality after cancer surgery: failure to rescue. JAMA surgery 2014; 149:475-81.

Oronce CI, Shao H, Shi L. Disparities in 30-Day Readmissions After Total Hip Arthroplasty. Medical care. 2015;53(11):924-930.

Singh JA, Lu X, Rosenthal GE, Ibrahim S, Cram P. Racial disparities in knee and hip total joint arthroplasty: an 18-year analysis of national Medicare data. Annals of the rheumatic diseases. 2014;73(12):2107-2115.

Skinner J, Chandra A, Staiger D, Lee J, McClellan M. Mortality after acute myocardial infarction in hospitals that disproportionately treat black patients. Circulation 2005; 112:2634-41.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

Below is a table showing the final variables in the model with associated odds ratios (OR).

Final Model Variables (variables meeting criteria in field 2b4.3) (**Dataset 1**)

Variable	07/2011-06/2014 OR (95% CI)
Age minus 65 (years above 65, continuous)	1.0 (1.03-1.04)
Male (%)	1.2 (1.14-1.18)
THA procedure	1.1 (1.07-1.12)
Number of procedures (two vs. one)	1.3 (1.27-1.43)
History of infection (CC 1, 3-6)	1.1 (1.07-1.12)
Metastatic cancer or acute leukemia (CC 7)	1.1 (0.99-1.25)
Cancer (CC 8-12)	1.0 (0.95-1)
Diabetes mellitus (DM) or DM complications (CC 15-20, 119, 120)	1.1 (1.11-1.16)
Protein-calorie malnutrition (CC 21)	1.3 (1.19-1.39)
Disorders of fluid/electrolyte/acid-base (CC 22-23)	1.1 (1.09-1.15)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 38)	1.2 (1.12-1.19)
Severe hematological disorders (CC 44)	1.4 (1.23-1.51)

Variable	07/2011-06/2014 OR (95% CI)
Dementia or other specified brain disorders (CC 49-50)	1.2 (1.15-1.24)
Major psychiatric disorders (CC 54-56)	1.3 (1.23-1.33)
Hemiplegia, paraplegia, paralysis, functional disability (CC 67-69, 100-102, 177-178)	1.2 (1.1-1.24)
Polyneuropathy (CC 71)	1.1 (1.11-1.19)
Congestive heart failure (CC 80)	1.2 (1.21-1.28)
Coronary atherosclerosis or angina (CC 83-84)	1.2 (1.17-1.23)
Hypertension (CC 89, 91)	1.3 (1.22-1.3)
Specified arrhythmias and other heart rhythm disorders (CC 92-93)	1.2 (1.13-1.18)
Stroke (CC 95-96)	1.1 (1.07-1.19)
Vascular or circulatory disease (CC 104-106)	1.1 (1.1-1.16)
Chronic obstructive pulmonary disease (COPD) (CC 108)	1.3 (1.3-1.37)
Pneumonia (CC 111-113)	1.2 (1.11-1.2)
Dialysis status (CC 130)	2.0 (1.72-2.25)
Renal failure (CC 131)	1.3 (1.23-1.3)
Decubitus ulcer or chronic skin ulcer (CC 148-149)	1.2 (1.12-1.24)
Cellulitis, local skin infection (CC 152)	1.1 (1.07-1.15)
Other injuries (CC 162)	1.1 (1.08-1.13)
Major symptoms, abnormalities (CC 166)	1.2 (1.15-1.2)
Morbid obesity (ICD-9 code 278.01)	1.3 (1.29-1.39)
Other congenital deformity of hip (joint) (ICD-9 code 755.63)	0.8 (0.62-1.03)
Post traumatic osteoarthritis (ICD-9 codes 716.15, 716.16)	1.0 (0.9-1.2)

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)

Variation in prevalence of the factor across measured entities

The prevalence of SES factors and African-American patients in the THA/TKA cohort varies across measured entities. The median percentage of dual-eligible patients is 6.7% (interquartile range [IQR] 3.9%- 11.8%). The median percentage of African-American patients is 2.1% (IQR 0.0%- 6.3%). The median percentage of patients with an AHRQ SES index score adjusted for cost of living at the census block group level equal to or below 42.7 is 12.9% (IQR 6.4%- 24.2%).

Empirical association with the outcome (univariate)

The patient-level observed (unadjusted) THA/TKA readmission rate is higher for dual-eligible patients, 6.7%, compared with 4.7% for all other patients. The readmission rate for African-American patients was also higher at 6.1% compared with 4.8% for patients of all other races. Similarly, the readmission rate for patients with an AHRQ SES index score equal to or below 42.7 was 5.5% compared with 4.8% for patients with an AHRQ SES index score above 42.7.

Incremental effect of SES variables and race in a multivariable model

We then examined the strength and significance of the SES and race variables in the context of a multivariable model. Consistent with the above findings, when we include any of these variables in a multivariable model that includes all of the claims-based clinical variables, the effect size of each of these variables is moderate (dual eligibility OR 1.22 [95% CI 1.18-1.27], race OR 1.19 [95% CI 1.14-1.24], AHRQ SES Index OR 1.09 [95% CI 1.06-1.12]). The c-statistic is virtually unchanged with the addition of any of these variables into the model. Furthermore, the addition of any of these variables into the model has little to no effect on hospital performance. We examined the change in hospitals' RSRRs with the addition of any of these variables. The median absolute change in hospitals' RSRRs when adding a dual eligibility indicator is 0.0199% (interquartile range [IQR] 0.0041% – 0.0418%, minimum -0.3252% – maximum 0.1133%), with a correlation coefficient between RSRRs for each hospital with and without dual eligibility added of 0.9982. The median absolute change in hospitals' RSRRs when adding a race indicator is 0.0217% (IQR 0.0037% – 0.0412%, minimum -0.6460% – maximum 0.0829%), with a correlation coefficient between RSRRs for each hospital with and without race added of 0.9990. The median absolute change in hospitals' RSRRs when adding an indicator for a low AHRQ SES index score adjusted for cost of living at the census block group level is 0.0172% (IQR 0.0048% – 0.0350%, minimum -0.0818% – maximum 0.1208%), with a correlation coefficient between RSRRs for each hospital with and without an indicator for a low AHRQ SES index score added of 0.9992.

Contextual Effect Analysis

As described in 2b4.3, we performed a decomposition analysis for each SES and race variable to assess whether there was a corresponding contextual effect. In order to better interpret the magnitude of results, we performed the same analysis for selected clinical risk factors. The results are described in the first table below (the decomposition table).

Both the patient-level and hospital-level effects for dual eligible,race, and low AHRQ SES Index were significantly associated with THA/TKA readmission in the decomposition analysis. **That the hospital-level effects were significant indicates that if the dual eligible or low AHRQ SES Index variables were used in the model to adjust for patient-level differences, then**

some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality.

To assess the relative contributions of the patient- and hospital-level effects, we calculated a range of predicted probabilities of readmission for the SES or race variables and clinical covariates (comorbidities), as described in section 2b4.3. The results are presented in the figure and second table below (table of predicted probabilities for SES and race variables).

For the AHRQ SES Index and race variables, the hospital-level effect (P95-P5) is greater than the patient-level effect (delta). For the dual eligibility variable, the patient-level effect is slightly greater than the hospital-level effect (second table below; the table of predicted probabilities for SES and race variables). Conversely, for clinical variables, the patient-level effect (delta) is greater than the hospital-level effect (P95-P5) for renal failure and COPD. For metastatic cancer, the patient-level effect is greater than the hospital-level effect, although neither effect is statistically significant likely due to small sample size (third table below; the table of predicted probabilities for clinical variables). This pattern demonstrates that the AHRQ SES index score and race variables have a much greater hospital-level effect than patient-level effect. However, the dual eligible status variable has a slightly higher patient-level effect. The clinical variables consistently had a greater effect at the patient level than at the hospital level. Therefore, including SES and race variables into the model would predominantly adjust for a hospital-level effect, potentially obscuring an important signal of hospital quality.

In the context of our conceptual model, we find clear evidence supporting the first two mechanisms by which SES might be related to poor outcomes. First we find that, although unadjusted rates of readmission are higher for patients of low SES or African-American race, the addition of SES to the readmission risk model, which already adjusts for clinical factors, makes very little difference. In particular, the odds ratios for SES and race variables are modest in the multivariate model, and there is little-to-no change in model performance or hospital results with the addition of SES. This suggests that the model already largely accounts for the differences in clinical risk factors (degree of illness and comorbidities) among patients of varied SES.

Second, the predominance of the hospital-level effect of SES and race variables in the decomposition analyses suggests that the risk associated with low SES is largely due to lower quality of care at hospitals where more patients with these risk factors are treated; hospitals caring for socially- and economically-disadvantaged patients have higher readmission risk for **all** of their patients. Patients with low SES or African-American race indicators tend to receive care more frequently at lower quality hospitals compared with patients with high SES indicators. Direct adjustment for patient SES would essentially "over-adjust" the measure. That is to say, it would be adjusting for an endogenous factor, one that influences the outcome through the site of treatment (hospital), as much as or more than through an attribute of the patient.

In comparison, we did not observe the same predominance of the hospital-level effect among the clinical covariates, reinforcing that SES and race factors have a causal pathway influenced by the hospital that is distinct from patient clinical attributes in their impact on readmission risk.

Summary

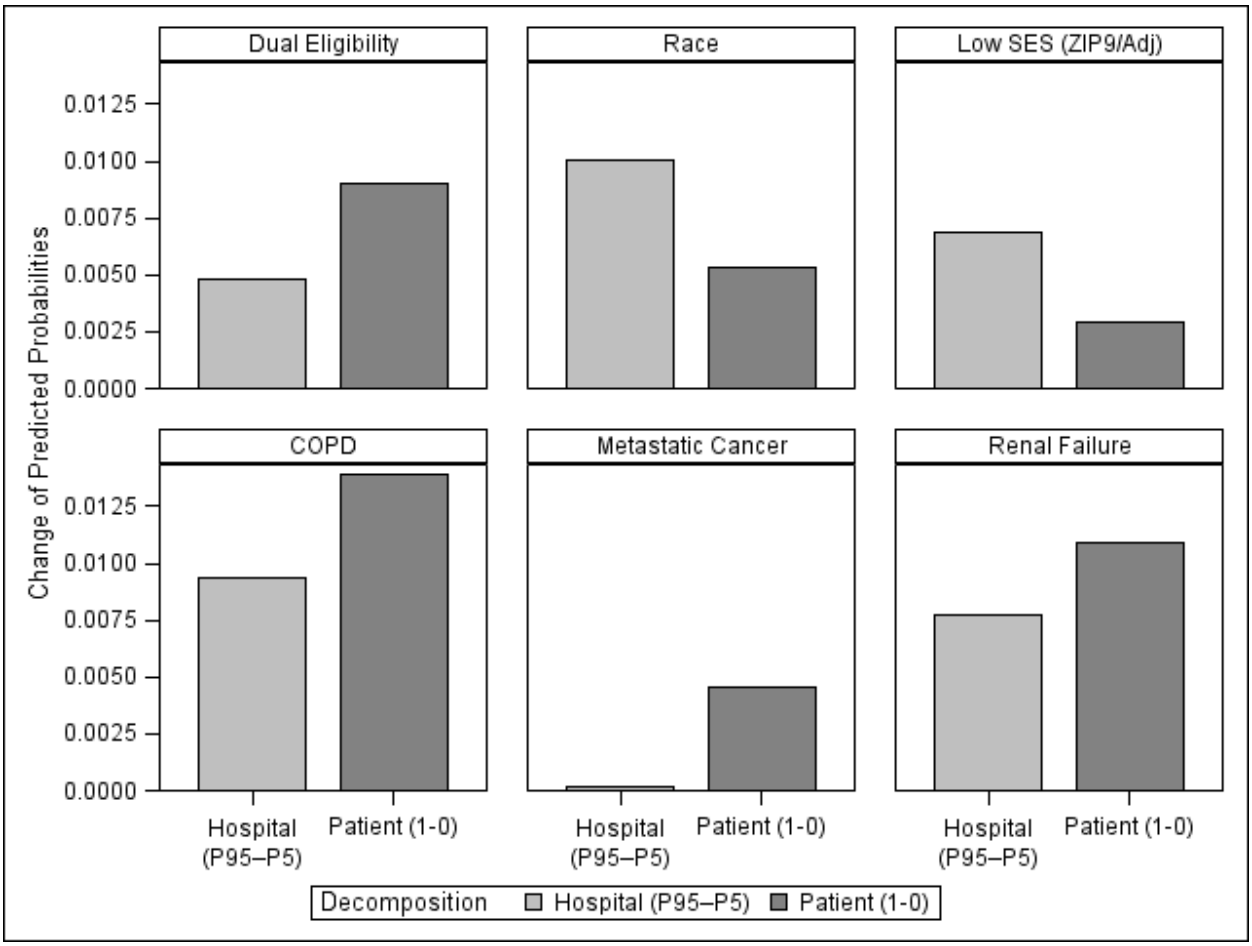
We found wide variation in the distribution of the three SES and race factors we examined, and we found that all three had some association with readmission risk. However, adjustment for these factors did not have an appreciable impact on hospital RSRRs, suggesting that existing

clinical risk factors capture much of the risk related to low SES. More importantly, we found that for the AHRQ SES index there was a greater hospital-level effect, compared with the patient-level effect and for dual eligible status the hospital effect was only slightly less than the patient effect, indicating that patient-level adjustment would adjust for quality differences between hospitals. Therefore, we did not include SES factors in our final risk model.

Hip/Knee Readmission Decomposition Analysis

Parameter	Estimate (Standard Error)	P-value
Dual Eligible – Patient-Level	0.1821 (0.0181)	<0.0001
Dual Eligible – Hospital-Level	0.3190 (0.0723)	<0.0001
African American – Patient-Level	0.1083 (0.0226)	<0.0001
African American – Hospital-Level	0.9248 (0.0817)	<0.0001
Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)– Patient-Level	0.0622 (0.0145)	<0.0001
Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)– Hospital-Level	0.3012 (0.0541)	<0.0001
Renal Failure – Patient-Level	0.2309 (0.0151)	<0.0001
Renal Failure – Hospital-Level	1.2792 (0.1903)	<0.0001
Metastatic Cancer – Patient-Level	0.1015 (0.0584)	0.0823
Metastatic Cancer – Hospital-Level	0.1960 (0.7556)	0.7953
COPD – Patient-Level	0.2770 (0.0128)	<0.0001
COPD – Hospital-Level	0.8837 (0.1144)	<0.0001

Change of Predicted Probabilities for SES and Race Compared with Clinical Variables in THA/TKA Readmission Measure



*Low SES (ZIP9/Adj) measured by linking patients' 9-digit ZIP codes to AHRQ SES Index at the census block group level, adjusted for cost of living

Predicted Probabilities for SES and Race Variables in the THA/TKA Readmission Measure

Hospital SES/Race Risk Factor Percentile	Dual Eligibility				African American Race				Low SES Census Block Group (AHRQ SES Index, Linked to 9-Digit ZIP – Adjusted for Cost of Living)			
	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)
0%	0.0000	0.0465	0.0551	0.0086	0.0000	0.0459	0.0508	0.0049	0.0000	0.0460	0.0487	0.0028
5%	0.0158	0.0467	0.0554	0.0086	0.0000	0.0459	0.0508	0.0049	0.0133	0.0461	0.0489	0.0028
10%	0.0233	0.0469	0.0555	0.0086	0.0000	0.0459	0.0508	0.0049	0.0267	0.0463	0.0491	0.0028
25%	0.0389	0.0471	0.0558	0.0087	0.0000	0.0459	0.0508	0.0049	0.0640	0.0468	0.0496	0.0028
50%	0.0670	0.0475	0.0562	0.0088	0.0206	0.0468	0.0517	0.0050	0.1295	0.0477	0.0505	0.0028
75%	0.1183	0.0482	0.0571	0.0089	0.0631	0.0485	0.0537	0.0051	0.2423	0.0492	0.0522	0.0029
90%	0.2121	0.0496	0.0587	0.0091	0.1402	0.0519	0.0573	0.0055	0.3931	0.0513	0.0544	0.0031
95%	0.3205	0.0512	0.0606	0.0094	0.2194	0.0555	0.0613	0.0058	0.4962	0.0528	0.0560	0.0031
100%	0.9375	0.0614	0.0725	0.0111	0.9404	0.1011	0.1110	0.0099	0.9565	0.0601	0.0636	0.0035
P95 – P5 (Hospital Effect)	-	0.0044	0.0052	-	-	0.0096	0.0105	-	-	0.0067	0.0071	-

Predicted Probabilities for Clinical Variables in the THA/TKA Readmission Measure

Hospital SES/Race Risk Factor Percentile	Renal Failure				Metastatic Cancer				Chronic Obstructive Pulmonary Disease			
	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)	Var_.J bar	Var_ij=0 for all patients	Var_ij=1 for all patients	Delta (Patient Effect)
0%	0.0000	0.0400	0.0497	0.0097	0.0000	0.0458	0.0504	0.0046	0.0000	0.0410	0.0531	0.0122
5%	0.0372	0.0419	0.0520	0.0101	0.0000	0.0458	0.0504	0.0046	0.0658	0.0433	0.0561	0.0128
10%	0.0482	0.0424	0.0527	0.0103	0.0000	0.0458	0.0504	0.0046	0.0824	0.0439	0.0568	0.0130
25%	0.0667	0.0434	0.0539	0.0105	0.0000	0.0458	0.0504	0.0046	0.1074	0.0448	0.0580	0.0132
50%	0.0879	0.0445	0.0552	0.0107	0.0034	0.0458	0.0504	0.0046	0.1406	0.0460	0.0596	0.0136
75%	0.1141	0.0459	0.0570	0.0111	0.0077	0.0459	0.0504	0.0046	0.1852	0.0478	0.0618	0.0140
90%	0.1452	0.0477	0.0591	0.0115	0.0133	0.0459	0.0505	0.0046	0.2353	0.0498	0.0644	0.0146
95%	0.1648	0.0488	0.0605	0.0117	0.0183	0.0460	0.0505	0.0046	0.2771	0.0515	0.0666	0.0150
100%	0.4054	0.0650	0.0801	0.0152	0.1200	0.0468	0.0515	0.0047	0.9431	0.0882	0.1124	0.0242
P95 – P5 (Hospital Effect)	-	0.0070	0.0085	-	-	0.0002	0.0002	-	-	0.0083	0.0105	-

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Approach to assessing model performance (**Dataset 1 and Dataset 2**)

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the cohorts:

Discrimination Statistics

(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic (also called ROC) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; good discrimination indicated by a wide range between the lowest decile and highest decile)

Calibration Statistics

(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients)

We tested the performance of the model for **Dataset 1** and **Dataset 2** described in section 1.7. During initial measure development, we tested the performance of the model developed in a randomly selected half of the admissions for THA/TKA in 2008, and compared performance calculated from admissions from the second half (**Dataset 2**). As a part of measure reevaluation, each year we assess temporal trends in model performance in the combined 3-year public reporting data (**Dataset 1**). Below, we report the model performance only for the 3-year combined results. For results for each individual year within the combined 3-year data, please see the attached specifications report.

Reference:

F.E. Harrell and Y.C.T. Shih, Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

For the development cohort (**Dataset 2**) the results are summarized below:

For the development sample the results are summarized below:

C-statistic= 0.65

Predictive ability (lowest decile %, highest decile %): 2.4%, 13.4%

For the validation sample the results are summarized below:

C statistic= 0.64

Predictive ability (lowest decile %, highest decile %): 2.6%, 13.2%

For the current measure cohort (**Dataset 1**) the results are summarized below:

C statistic = 0.65; Predictive ability (lowest decile %, highest decile %) = (1.8, 10.9)

For comparison of model with and without inclusion of SDS factors, see Section 2b4.4b.

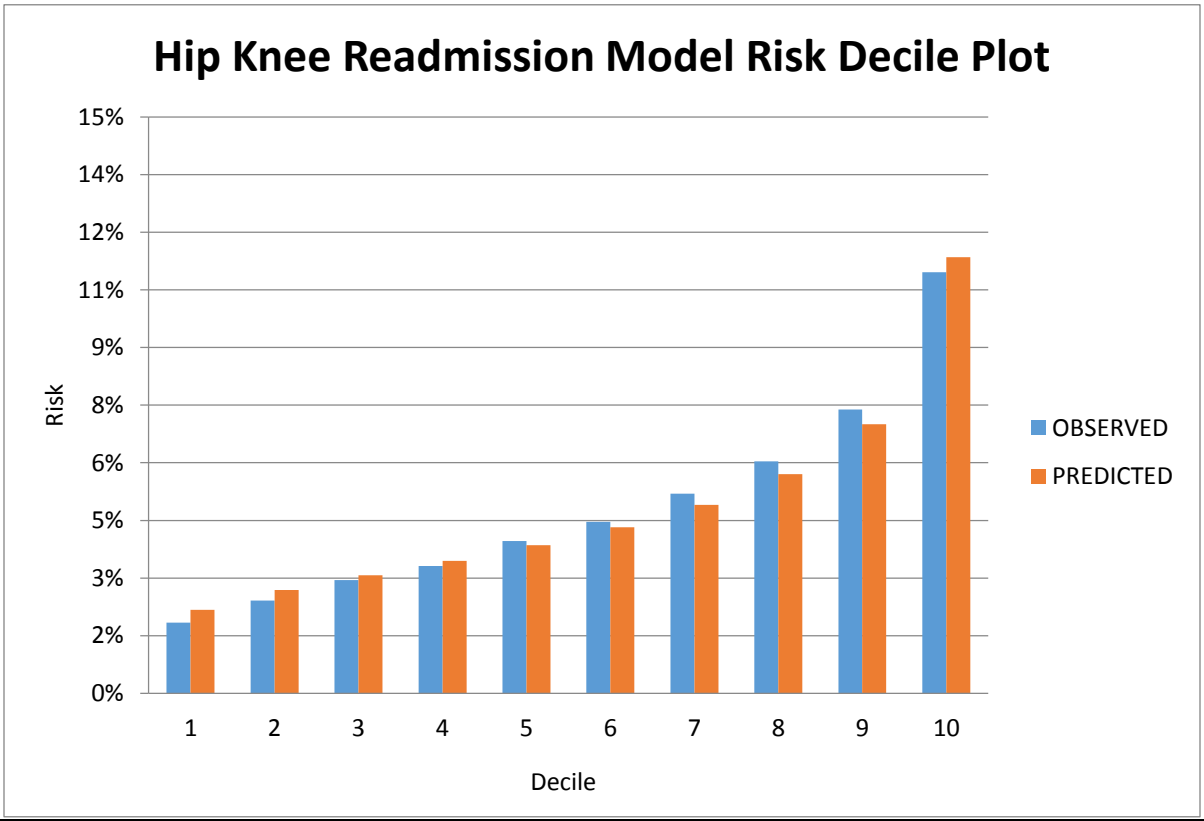
2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

For the original measure development cohort (**Dataset 2**) the results are summarized below:

First half of split sample: Calibration: (0, 1)
Second half of split sample: Calibration: (-0.06, 0.98)

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from July 2011 to June 2014 (**Dataset 1**).



2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.e., *what do the results mean and what are the norms for the test conducted*)

Discrimination Statistics

The C-statistics of 0.65, 0.64, and 0.65 for the model development, validation, and current public reporting data (**Datasets 2, 2,** and **1** respectively) demonstrate consistent and fair model discrimination (. The models also

indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

Calibration Statistics

Over-fitting (Calibration γ0, γ1)

If the γ0 in the development and validation samples (**Dataset 2**) are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration value of close to 0 at one end (first half of split sample) and close to 1 to the other end (second half of split sample) indicates good calibration of the model.

Risk Decile Plots

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates excellent discrimination of the model and good predictive ability.

Overall Interpretation

Interpreted together, our diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

Application to Medicare FFS Beneficiaries Using Inpatient Data Only for Risk Adjustment

In testing other administrative claims measures developed in Medicare FFS data – including mortality and readmission measures for acute myocardial infarction (AMI), heart failure, pneumonia, and chronic obstructive pulmonary disease (COPD) – we have validated both the accuracy of the California Patient Discharge Data (PDD) in capturing Medicare claims and the use of only inpatient data for risk adjustment[1, 2]. We also found that, although the prevalence of most risk factors is lower when using only inpatient claims data, the magnitude of effect for most risk factors was similar when comparing the models that use all patient history data and those that use only inpatient claims data. Over 95% of patients were in a similar risk category (defined as being in the same or adjacent category) regardless of the risk-adjustment dataset used, and the integrated discrimination improvement values were relatively low (ranging from -0.001 for COPD readmission, to 0.005 for COPD mortality). For all measures, the c-statistic was also qualitatively similar between the two approaches (the greatest difference in c- statistic between inpatient only versus all patient history data risk-adjustment models was 0.012, for AMI mortality). Moreover, when comparing the models using full history data with the models using only inpatient claims data; hospital-level risk-standardized rates were highly correlated (ICCs ranged from 0.978 for COPD readmission to 0.986 for COPD mortality). Based on this reassuring data across measures and the fact that we anticipate key comorbidities for an all-payer population of THA/TKA patients, such as rheumatoid arthritis, will similarly be captured by inpatient coding, we did not repeat these analyses for the THA/TKA readmission measure, but rather assumed that inpatient claims data would provide adequate risk-adjustment information for application of the measure in all-payer data.

Testing of Measure in All-Payer Cohort

We tested the THA/TKA readmission measure in an all-payer patient population of adults aged 18 years and older so that it can be applied to both Medicare and all-payer populations. Using data from California, as well as CMS Medicare FFS data for California hospitals, we performed analyses to determine whether the THA/TKA readmission measure can be applied to all adult patients, including FFS Medicare patients aged 65+, non-FFS Medicare patients aged 65+, and patients aged 18-64 years at the time of admission.

To address the question of how well the models perform when applied to all patients 18+, we used the California Patient Discharge Data (PDD). Specifically, using 2006 data, we created measure cohorts with up to one year of hospital inpatient claims history and 30-day follow-up data. For the THA/TKA readmission measure, we:

1. Created the patient cohorts using the respective measure inclusion and exclusion criteria (with the exception of including all patients 18+), and compared the FFS 65+, non-FFS 65+, and 18-64 year-old patient subgroups with respect to the distributions of risk factors and the crude outcome rates.

2. Fit the models in all patients 18+ and: (i) examined overall model performance in terms of the C statistic, (ii) compared performance (C statistic and predictive ability) across the patient subgroups (FFS 65+, non-FFS 65+, all 65+, and all-payer 18-64), and (iii) compared the distribution of Pearson residuals (model fit) across the patient subgroups.

3. Fit the models separately in each patient subgroup and compared odds ratios (ORs) associated with the risk factors to assess differences in magnitude or direction of ORs among the subgroups.

To determine whether the relationship between each risk factor and the outcome differed for those aged 65+ vs. 18-64 in ways that would affect measure results, we:

1. Fit the models in all patients 18+ and tested interaction terms between age (65+ vs. 18-64) and each of the other risk factors.

2. Fit the models in all patients 18+ with interaction terms and compared performance (C statistic and predictive ability) across the patient subgroups.

3. Fit the models in all patients 18+ with and without interaction terms and (i) conducted a reclassification analysis to compare risk prediction at the patient level; (ii) compared the C statistic; and (iii) compared hospital-level risk-standardized rates using a scatterplot and the ICC to assess whether the models with interactions are statistically different from the current models in profiling hospital rates.

All patient-level models were estimated using a logistic regression model; next, hospital-level RSRR analyses were conducted using a hierarchical logistic regression model approach.

Results of Measure Testing in All-Payer Cohort

There are some differences in the risk factor profiles and crude outcome rate among patient subgroups. In general, the prevalence of risk factors was similar in FFS 65+ and non-FFS 65+ patients. There were slight differences in the prevalence of two risk factors in the readmission measure (Chronic Atherosclerosis [CC 83-84] and Hypertension [CC 89, 91]). When comparing risk factor prevalence estimates between those 65+ and younger patients aged 18-64, frequencies were generally either lower in the younger cohort or similar between the groups. For some risk factors, including having an index elective THA procedure and morbid obesity (ICD-9 code 278.01) in the readmission model, prevalence estimates were in fact higher in younger than in older patients. The readmission rates were very similar between older patients 65+ and younger patients 18-64 years.

When the current models were applied to all patients 18+, overall discrimination was good (C statistic= 0.64 for THA/TKA readmission). There was also good discrimination and predictive ability in all subgroups of patients. Moreover, for both measures, the distribution of Pearson residuals was comparable across the patient subgroups.

ORs were generally similar for FFS 65+ and non-FFS 65+ patients. For some risk factors, such as Post-traumatic osteoarthritis (ICD-9 codes 716.15, 716.16), there were differences in the magnitude of effect between younger and older patients.

There were six significant age-by-risk-factor interaction terms for readmission (Older and History of Infection [CC 1, 3-6], Older and Diabetes and DM Complications [CC 15-20, 119, 120], Older and Disorders of fluid/electrolyte/acid-base [CC 22, 23], Older and COPD [CC 108], Older and Major Symptoms, abnormalities [CC 166], Older and Morbid obesity [ICD-9 code 278.01]). Inclusion of the interaction terms, however, did not substantively change the level of discrimination and predictive ability across the patient subgroups.

In addition, when comparing patient risk classifications for the measure with and without interaction terms, the reclassification analysis demonstrated good patient-level risk prediction: for all patient subgroups, nearly 100% of patients were in a similar risk category (defined as being in the same or adjacent category), regardless of risk-adjustment strategy. Moreover, the C-statistic was nearly identical for the models with and without interaction terms for THA/TKA readmission (0.63 vs. 0.64, respectively). Finally, when comparing each measure with and without interaction terms, the hospital-level risk-standardized rates estimated by the two versions of each model were highly correlated (r = 0.998 for both THA/TKA readmission).

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

For public reporting of the measure, CMS characterizes the uncertainty associated with the RSRR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSRR's interval estimate does not include the national observed readmission rate (is lower or higher than the rate), then CMS is confident that the hospital's RSRR is different from the national rate, and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate," respectively. If the interval includes the national rate, then CMS describes the hospital's RSRR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Analyses of Medicare FFS data show substantial variation in RSRRs among hospitals. Using data from July 2011 – June 2014 (**Dataset 1**), the median hospital RSRR was 4.8%, with a range of 2.6% to 8.6%. The interquartile range was 4.6%-5.2%.

Out of 3,498 hospitals in the U.S., 49 performed "better than the U.S. national rate," 2,721 performed "no different from the U.S. national rate," and 49 performed "worse than the U.S. national rate." 679 were classified as "number of cases too small" (fewer than 25) to reliably tell how well the hospital is performing.

Note that this analysis included index admissions from July 2011 – June 2014 from the 2015 public reported data (**Dataset 1**). Please note that these analyses were done using version 3.0 of the planned readmission algorithm rather than the current version 4.0.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The variation in rates and number of performance outliers suggests that differences remain in the quality of care received across hospitals for THA/TKA. This evidence supports continued measurement to reduce the variation.

Note: From the July 2011 to June 2012 reporting year to the July 2013 to June 2014 reporting year, the observed THA/TKA readmission rate decreased from 5.3% (July 2011 – June 2012) to 4.5% (July 2013 – June 2014). The observed readmission rate for the 3-year combined public reporting period (July 2011 – June 2014) for THA/TKA Medicare FFS patients is 4.9%.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS 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.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

N/A

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

N/A

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic claims

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those

whose performance is being measured.
Administrative data are routinely collected as part of the billing process.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
There are no fees associated with the use of this measure.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Not in use	Public Reporting Hospital Inpatient Quality Reporting (IQR) Program http://cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalRHQDAPU.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Hospital Readmission Reduction (HRRP) Program http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Public Reporting
Program Name, Sponsor: Hospital Inpatient Quality Reporting (Hospital IQR) Program, Centers for Medicare and Medicaid Services (CMS)

Purpose: The Hosptial IQR program was originally mandated by Section 501(b) of the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003. This section of the MMA authorized CMS to pay hospitals that successfully report designated quality measures a higher annual update to their payment rates. Initially, the MMA provided for a 0.4 percentage point reduction in the annual market basket (the measure of inflation in costs of goods and services used by hospitals in treating Medicare patients) update for hospitals that did not successfully report. The Deficit Reduction Act of 2005 increased that reduction to 2.0 percentage points.

In addition to giving hospitals a financial incentive to report the quality of their services, the Hospital IQR program provides CMS with data to help consumers make more informed decisions about their health care. Some of the hospital quality of care information gathered through the program is available to consumers on the Hospital Compare website at: www.hospitalcompare.hhs.gov.

Geographic area and number and percentage of accountable entities and patients included:

The IQR program includes all Inpatient Prospective Payment System (IPPS) non-federal acute care hospitals and VA hospitals in the United States. The number and percentage of accountable hospitals included in the program, as well as the number of patients included in the measure, varies by reporting year. For 2015 public reporting, the RSRR was reported for 4,663 hospitals across the U.S. The final index cohort includes 925,315 admissions.

Payment Program
Program Name, Sponsor: Hospital Readmission Reduction (HRRP) Program, Centers for Medicare and Medicaid Services (CMS)

Purpose: Section 3025 of the Affordable Care Act added section 1886(q) to the Social Security Act establishing the Hospital Readmissions Reduction Program, which requires CMS to reduce payments to IPPS hospitals with excess readmissions, effective for discharges beginning on October 1, 2012. The regulations that implement this provision are in subpart I of 42 CFR part 412 (§412.150 through §412.154).

Geographic area and number and percentage of accountable entities and patients included: The HRRP program includes only Subsection (d) hospitals and hospitals located in Maryland. Subsection (d) hospital encompasses any acute care hospital located in one of the fifty States or the District of Columbia which does not meet any of the following exclusion criteria as defined by the Social Security Act: psychiatric, rehabilitation, children's, or long-term care hospitals, and cancer specialty centers. By definition, all other hospitals are considered subsection (d) hospitals. This means that critical access hospitals, cancer hospitals, and hospitals located in U.S territories will not be included in the calculation. The number and percentage of accountable entities included in the program, as well as the number of patients included in the measure, varies by reporting year.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
N/A. This measure is currently publicly reported.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
N/A. This measure is currently publicly reported.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

There has been progress in 30-day RSRR for THA and/or TKA. The median 30-day RSRR decreased by 0.8 absolute percentage points from July 2011-June 2012 (median RSRR: 5.2%) to July 2013-June 2014 (median RSRR: 4.4%). The median hospital RSRR from July 2012-June 2014 was 4.8% (IQR 4.6% - 5.2%).

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We did not identify any unintended consequences during measure development, model testing, or re-specification. However, we are committed to monitoring this measure's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care, increased patient morbidity and mortality, and other negative unintended consequences for patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0330 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following heart failure (HF) hospitalization
0505 : Hospital 30-day all-cause risk-standardized readmission rate (RSRR) following acute myocardial infarction (AMI) hospitalization.
0506 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following pneumonia hospitalization
1550 : Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
1789 : Hospital-Wide All-Cause Unplanned Readmission Measure (HWR)
1891 : Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization
The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
We did not include in our list of related measures any non-outcome measures (for example, process measures) with the same target population as our measure. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

5b. Competing Measures
The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed**

measure(s):
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** ProcSpecific_Rdmsn_Rpt_2016.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** Centers for Medicare & Medicaid Services
Co.2 **Point of Contact:** Lein, Han, Lein.han@cms.hhs.gov, 410-786-0205-
Co.3 **Measure Developer if different from Measure Steward:** Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/CORE)
Co.4 **Point of Contact:** Karen, Dorsey, karen.dorsey@yale.edu, 203-764-5700-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The working group involved in the initial measure development is detailed in the original technical report available at www.qualitynet.org.

Our measure development team consisted of the following members:
Laura M. Grosso, PhD, MPH
Jeptha P. Curtis, MD
Zhenqiu Lin, PhD
Lori L. Geary, MPH
Smitha Vellanky, MSc
Carol Oladele, MPH
Yongfei Wang, MS
Elizabeth E. Drye, MD, SM
Harlan M. Krumholz, MD, SM

Workgroup Members:
Daniel J. Berry, MD
Kevin J. Bozic, MD, MBA
Robert Bucholz, MD
Lisa Gale Suter, MD
Charles M. Turkelson, PhD
Lawrence Weis, MD

Technical Expert Panel Members:
Mark L. Francis, MD
Cynthia Jacelon, PhD, RN, CRRN
Norman Johanson, MD
C. Kent Kwoh, MD
Courtland G. Lewis, MD
Jay Lieberman, MD

Peter Lindenauer, MD, M.Sc.
Russell Robbins, MD, MBA
Barbara Schaffer
Nelson SooHoo, MD, MPH
Steven H. Stern, MD
Richard E. White, Jr., MD

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2013

Ad.3 **Month and Year of most recent revision:** 06, 2015

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3017
De.2. **Measure Title:** PBM-02: Preoperative Hemoglobin Level
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure is designed to allow transfusion/blood use review committees to identify patients undergoing elective surgery with suboptimal, uncorrected hemoglobin levels that may have led to perioperative transfusion. This measure assesses, via stratification, pre-operative hemoglobin levels of selected elective surgical patients age 18 and over who received a perioperative red blood cell transfusion.
1b.1. **Developer Rationale:** There are many corrective interventions available for patients identified with preoperative sub-optimal hemoglobin levels in order to avoid a transfusion during or after the surgical procedure. As an essential component of blood management, pre-operative investigation and correction of anemia should be undertaken, since transfusion has been shown to increase adverse outcomes. Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need5.

One study of hip and knee arthroplasty patients found that those with a hemoglobin level <13.0g/dL. had four times the risk for blood transfusion than those with higher hemoglobin levels5.

Prevalence of preoperative anemia varies by population: Community-dwelling, >65 years old - <10%
i. Frail nursing home resident - >48%
ii. Surgical population – 5% to 75%
iii. Octogenarian, elective cardiac surgery – 49.4%1
iv. 7% of 9,462 patients undergoing total hip or total knee replacement2
v. >65 years old – 11% women, 10.2% men (NHANES Study)3
vi. Elective orthopedic surgery – 35%4

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

Preoperative anemia is also a predictor of postoperative transfusion in orthopedic, major colon, and major cardiac surgery. Since blood transfusion is the most frequently-performed hospital procedure (11% of hospital stays) and has increased by 126% from 1997 – 2010, and since blood transfusion can have adverse outcomes, such as prolonged length of stay and decreased functional status at discharge, investigation and correction of preoperative anemia is essential to any blood management program.

The World Health Organization has defined the levels of anemia for men at a hemoglobin measurement of less than 13.0, and for non-pregnant women at a hemoglobin measurement of less than 12.0. There has, however, been controversy over these levels. While there is debate regarding the hemoglobin level at which patients are considered anemic7, use of the WHO definition of anemia allows identification of patients for whom pre-operative investigation and correction of hemoglobin levels is warranted.

The intent of the measure is to provide information to providers and review groups about the incidence of transfusions in the various

strata, with the objective of identifying trends related to over- and underutilization of blood transfusions and correction of preoperative anemia.

5. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
6. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. The Journal of Bone and Joint Surgery. Volume 84-A – Number2 – February 2002.
7. Beutler E, Waalen J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? Blood Mar 1 2006 (107)5: 1747-1750.

S.4. **Numerator Statement:** Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
12.0 grams or above
>=11.0 and <12.0 grams (mild anemia)
>=8.0 and <11.0 grams (moderate anemia)
Below 8.0 grams (severe anemia)

S.7. **Denominator Statement:** Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.

S.10. **Denominator Exclusions:**
- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
- Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure.
- Patients with sickle cell disease or hereditary hemoglobinopathy

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

The developer provides the following path to support the relationship between the process of care (optimized hemoglobin levels prior to elective surgery) and outcomes:

1. Process: Optimized preoperative hemoglobin level

2. Elective surgical procedure performed
3. Reduced rate of blood transfusion
4. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

- The developer provided a guideline from the Network for Advancement of Transfusion Alternatives:
  - *Recommendation 2*: *We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl^{-1}, male ≥13 g dl^{-1}), according to the WHO criteria (Grade 2C). This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anaemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.* Grade 2C: Weak recommendation ("we suggest") and low or very low quality evidence (observational studies, randomized controlled tried with major limitations).
- The developer provided an additional 15 citations as sources of evidence for this measure, many of which were surgery type specific. Conclusions in the Fowler AJ et al. article published in 2015 state: "*Preoperative anaemia is associated with poor outcomes after surgery, although heterogeneity between studies was significant. It remains unclear whether anaemia is an independent risk factor for poor outcome or simply a marker of underlying chronic disease. However, red cell transfusion is much more frequent amongst anaemic patients*."

Guidance from the Evidence Algorithm

Based on SR/grading of clinical practice guideline (Box 3) → QQC provided (Box 4) → Moderate quality evidence based on additional relevant review articles provided (Box 5b) → MODERATE

Questions for the Committee:

- *Is the evidence directly applicable to the determination of preoperative hemoglobin level and the reduced risk of transfusion-related adverse outcomes?*
- *How strong is the evidence for this relationship?*

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Performance Gap

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Although there is no performance data on the measure as specified, the developer listed data with citations from the literature that indicates opportunity for improvement that relate to the focus of measurement.

Disparities

- The developed indicated that no disparity data are available.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support the Measure Focus

- Decreasing blood use is worthy
  This measure will show a large gap, but there is noise with the signal. patients with renal failure, for instance.
  Seems redundant with previous measure -- would be OK with one or the other, but not both
  Measure title not on list? What is 3017? Text is for Preop HGB testing; 3017 title is blood group testing

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
  - 12.0 grams or above
  - >=11.0 and <12.0 grams (mild anemia)
  - >=8.0 and <11.0 grams (moderate anemia)
  - Below 8.0 grams (severe anemia)
- Denominator Statement: Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.
- Denominator Exclusions:
  - Patients under age 18
  - Patients whose surgical procedure is performed to address a traumatic injury
  - Patients who have a solid organ transplant
  - Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
  - Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure
  - Patients with sickle cell disease or hereditary hemoglobinopathy
- Level of Analysis: Facility
- Care Setting: Hospital/Acute Care Facility
- No risk adjustment or risk stratification

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM;

Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously; **Bonnie results submitted**
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 78 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications, including the hemoglobin level thresholds, are consistent with evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.38
Denominator clearly describes the activity being measured	4.46

Numerator inclusions clear and appropriate	4.51
Denominator inclusions clear and appropriate	4.53
Numerator exclusions clear and appropriate	4.44
Denominator exclusions clear and appropriate	4.45
Accurately assesses the process of care to which it is addressed	4.13

This measure is being considered for trial use, thus full validity testing results are not expected and the Committee will not vote on this criterion.

2b3-2b7. Threats to Validity

2b3. Exclusions:
When data are available, the developer plans to analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Maternal and Fetal procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- ECMO procedures recorded in SNOMEDCT or ICD10PCS that start prior to the elective surgical procedure
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing any of the following conditions:
  - Traumatic Injury
  - Pregnancy, Childbirth, and the Puerperium
- Sickle Cell Disease and Related Blood disorders

Questions for the Committee:
- *Are there other threats to validity the measure developer should consider?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time

Question for the Committee:
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data
The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms None identified

Feedback :

None identified

***Questions for the Committee**:*

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

•

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: PBM-02: Preoperative Hemoglobin Level

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.

- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Optimized hemoglobin levels prior to elective surgery.

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

5. Process: Optimized preoperative hemoglobin level
6. Elective surgical procedure performed
7. Reduced rate of blood transfusion
8. Outcomes: A. Reduced risk of transfusion-related adverse outcomes, which can include decreased functional status at discharge, prolonged length of stay, increased mortality, and complications of transfusion, such as TRALI, hemolytic reactions, and other incompatibilities/complications. B. Reduced resource (blood) usage.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Goodnough LT, Maniatis A, Earnshaw P, Benoni G, et al. Detection evaluation, and management of preoperative anaemia in the elective orthopaedic surgical patient: NATA Guidelines. *Br. Journ. Anesthesia*, 106 (1): 13-22 (2011).

http://bja.oxfordjournals.org/content/106/1/13.full

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 2: We suggest that the patient's target Hb before elective surgery be within the normal range (female ≥12 g dl^{-1}, male ≥13 g dl^{-1}), according to the WHO criteria (Grade 2C).

This recommendation is a suggestion, indicating a lack of panel consensus and evidence on whether elective surgical procedures should be cancelled, representing best practices, for patients who are identified to be anaemic. Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

Grade 2C

Grading system

Strength of recommendation: is risk/benefit clear?

- Yes ⇨ strong recommendation=Grade 1: 'we recommend'
- No ⇨ weak recommendation=Grade 2: 'we suggest'

Quality of evidence

- High-quality evidence=A (meta-analyses, randomized controlled trials)
- Moderate-quality evidence=B (randomized controlled trials with limitations, observational studies with large effects)
- Low- or very low-quality evidence=C (obervational studies, randomized controlled tried with major limitations)

Grade of recommendation=6 possible grades

- Grade 1A
- Grade 1B
- Grade 1C
- Grade 2A
- Grade 2B
- Grade 2C

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

See above

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same as 1a.4.1

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Detection, evaluation, and management of preoperative anemia in elective orthopedic surgery.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Grade C – low-quality evidence (observational studies, randomized control trials with major limitations).

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

See 1a.4.3

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: 1966 – January 2010

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

5 observational studies, 3 cohort studies, 1 meta-analysis, 1 systematic literature review.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Not stated in citation; in review of studies appears that 3 are small cohort studies.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Unstated in citation

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

Delay of elective scheduled surgery for definitive evaluation of newly detected anaemia and associated clinical conditions (nutritional deficiency, chronic renal disease, etc.) will benefit patients and reduce harm, including likelihood of exposure to blood transfusions.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

None

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, MEDLINE and other relevant sources including professional association websites, The Cochrane Library, the National Guideline Clearinghouse, and other sources was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

1. American Red Cross: "Preoperative assessment and efforts to reduce the RBC transfusion requirement in the perioperative period include the evaluation and treatment of anemia prior to surgery and the evaluation for discontinuation or replacement of anticoagulant and antiplatelet medications …for a sufficient time prior to surgery in consultation with the prescribing physician." A Compendium of Transfusion Practice Guidelines, Second Edition, 2013, page 13.
2. Society for Blood Management: The panel further recommended that the patient's target hemoglobin before elective surgery should be within the normal range (normal female >12 g/dL, normal male >13 g/dL). Goodnough LT, Shander A, Spivak JL, Waters JH, et al. Detection Evaluation, and Management of Anemia in the Elective Surgical Patient. *Anesth Analg* 2005;101:1858-61, p. 1860
3. "Preoperative anaemia is associated with poor outcomes after surgery" Fowler AJ et al. Meta-analysis of the association between preoperative anaemia and mortality after surgery. *Br J Surg* 2015 Oct;102(11):1314-24.

METHODS:

A systematic review and meta-analysis of observational studies exploring associations between preoperative anaemia and postoperative outcomes was performed. Studies investigating trauma, burns, transplant, paediatric and obstetric populations were excluded. The primary outcome was 30-day or in-hospital mortality. Secondary outcomes were acute kidney injury, stroke and myocardial infarction. Predefined analyses were performed for the cardiac and non-cardiac surgery subgroups. A post hoc analysis was undertaken to evaluate the relationship between anaemia and infection. Data are presented as odds ratios (ORs) with 95 per cent c.i.

RESULTS:

From 8973 records, 24 eligible studies including 949 445 patients were identified. Some 371 594 patients (39·1 per cent) were anaemic. Anaemia was associated with increased mortality (OR 2·90, 2·30 to 3·68; I(2) = 97 per cent; P < 0·001), acute kidney injury (OR 3·75, 2·95 to 4·76; I(2) = 60 per cent; P < 0·001) and infection (OR 1·93, 1·17 to 3·18; I(2) = 99 per cent; P = 0·01). Among cardiac surgical patients, anaemia was associated with stroke (OR 1·28, 1·06 to 1·55; I(2) = 0 per cent; P = 0·009) but not myocardial infarction (OR 1·11, 0·68 to 1·82; I(2) = 13 per cent; P = 0·67). Anaemia was associated with an increased incidence of red cell transfusion (OR 5·04, 4·12 to 6·17; I(2) = 96 per cent; P < 0·001). Similar findings were observed in the cardiac and non-cardiac subgroups.

CONCLUSION:

Preoperative anaemia is associated with poor outcomes after surgery, although heterogeneity between studies was significant. It remains unclear whether anaemia is an independent risk factor for poor outcome or simply a

marker of underlying chronic disease. However, red cell transfusion is much more frequent amongst anaemic patients.

4. British Committee for Standards in Haemotology: Recommendation: "Healthcare pathways should be structured to ensure anaemia screening and correction before surgery." Kotze A, Harris A, Baker C, Iqbal T, eta l. British Committee for Standards in Haemotology Guidelines on the Identification and Management of Pre-Operative Anemia. *British Journal of Haemotology* Volume 171, Issue 3: November 2015 pages 322-331.

5. Society for the Advancement of Blood Management: "Patients who are having a procedure for which preoperative screening is required are identified at least three to four weeks prior to surgery to allow sufficient time to diagnose and manage anemia, unless the surgery is of an urgent nature and must be performed sooner." (Standard 6.2) SABM Administrative and Clinical Standards for Patient Blood Management Programs, Third Edition. Unpublished work, 2014. Downloaded from www.SABM.org on April 9, 2016.

6. New York State Department of Health: "Careful evaluation of pre-existing anemia and its treatment prior to surgery are an effective strategy for reducing surgical transfusion requirements." New York State Council on Human Blood and Transfusion Services. Guidelines for Transfusion Options and Alternatives, 2010. Downloaded from www.wadswoth.org/labcert/blood_tissue July 2015.

7. 13 references (12 articles, one literature review) document increased rate of perioperative blood transfusion when preoperative anemia is present. Ferraris et al., "Perioperative Blood Transfusion and Blood Conservation in Cardiac Surgery: The Society of Thoracic Surgeons and The Society of Cardiovascular Anesthesiologists Clinical Practice Guideline". *Ann Thorac Surg* 2007;83: 527 – 86.

8. 1 study of 296 elective orthopedic surgeries indicated through multivariate analysis that a significant relationship existed only between the need for transfusion and the preoperative hemoglobin level (p+ 0.00001) after hip and knee replacement. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. *The Journal of Bone and Joint Surgery*. Volume 84-A – Number2 – February 2002.

9. 1 systematic literature review of 29 included citations demonstrated that low hemoglobin and patient age were consistent risk factors for blood transfusion in orthopedic surgery. Barr PJ et al. Drivers of Transfusion Decision Making and Quality of the Evidence in Orthopedic Surgery: A Systematic Review of the Literature. *Transfusion Medicine Reviews*, Vol 25 No. 4 (October), 2011 pp. 304 – 316.
10. In a cohort study of 239 patients scheduled for transcatheter aortic valve implantation (TAVI), 62.3% were found to be anemic pre-procedurally and were referred to a blood conservation clinic (BCC) where they received a regimen of IV iron, oral iron, or epoetin alfa. Rates of transfusion in this cohort of 60 patients were assessed and compared with transfusion rates for TAVI patients prior to the initiation of the program. Implementation of the BCC was associated with a substantial decrease in the average blood transfusion rate from 33.3% before program initiation to 15.3% after implementation ($P < 0.001$). After adjusting for baseline hemoglobin values and comorbidities, being assessed at the BCC was strongly associated with a reduction in the need for transfusion (odds ratio, 0.28; 95% confidence interval, 0.11-0.69; P ¼ 0.006. Shuvy M, et al. Preprocedure Anemia Management Decreases Transfusion Rates in Patients Undergoing Transcatheter Aortic Valve Implantation. *Canadian Journal of Cardiology* (2016) Article in press.

11. A placebo-controlled, double-blind trial enrolling 316 patients scheduled for major, elective orthopedic hip or knee surgery who were expected to require 2.2 units of blood and who were not able or willing to participate in an autologous blood donation program examined the efficacy of Epogen treatment in reducing use of perioperative blood transfusion. Based on previous studies which demonstrated that pretreatment hemoglobin is a predictor of risk of receiving transfusion, patients were stratified into one of three groups based

on their pretreatment hemoglobin [-< 10 (n = 2) > 10 to 5 13 (n = 96), and > 13 to I 15 g/dL (n = 218)] and then randomly assigned to receive 300 Units/kg EPOGENQ 100 Units/kg EPOGEN@ or placebo by SC injection for 10 days before surgery, on the day of surgery, and for 4 days after surgery. All patients received oral iron and a low-dose post-operative warfarin. Treatment with EPOGENB 300 Units/kg significantly (p = 0.024) reduced the risk of allogeneic transfusion in patients with a pretreatment hemoglobin of > 10 to _< 13 g/dL; 5/31 (16%) of EPOGENB 300 Units/kg, 6126 (23%) of EPOGEN@ 100 Units/kg, and 13/29 (45%) of placebo treated patients were transfused. There was no significant difference in the number of patients transfused between EPOGENB (9% 300 Units/kg, 6% 100 Units/kg) and placebo (13%) in the > ***13*** to I ***15*** g/dL hemoglobin stratum. There were too few patients in the I 10 g/dL group to determine if EPOGEN@ is useful in this hemoglobin strata. In the > 10 to I 13 g/dL pretreatment stratum, the mean number of units transfused per EPOGENQ-treated patient (0.45 units blood for 300 Units/kg, 0.42 units blood for 100 Units/kg) was less than the mean transfused per placebo-treated patient (1.14 units) (overall p = 0.028). In addition, mean hemoglobin, hematocrit and reticulocyte counts increased significantly during the pre-surgery period in patients treated with EPOGEN. deAndrade JH, Jove M. Baseline Hemoglobin as a Predictor of Risk of Transfusion and Response to Epoetin alfa in Orthopedic Surgical Patients. *Am J of Orthoped*. 1996;25(8): 533-542.

12. Among 569 patients who underwent colorectal cancer surgery between 1998 and 2003, 32 anemic patients who received iron supplementation for at least 2 weeks preoperatively (group A) and 84 anemic patients who did not (group B) were studied.

There were no significant differences between groups A and B in age, sex, surgical technique, tumor stage, and operating time. Their Hgb and Hct values were similar at first presentation, but significantly different immediately before surgery (both P < 0.0001). There were no significant differences in intraoperative blood loss between the groups, but significantly fewer patients in group A needed an intraoperative blood transfusion (9.4% vs 27.4%, P < 0.05). Okuyama M et al. Preoperative iron supplementation and intraoperative transfusion during colorectal cancer surgery. *Surg Today*. 2005;35(1):36-40.

13. 1 systematic literature review of 13 studies including >29,000 orthopedic surgical patients showed that

a. The prevalence of preoperative anemia was 21-56%

b. Perioperative anemia was associated with an elevated blood transfusion rate, postoperative infections, poorer physical functioning and recovery, increased length of stay and mortality.

Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. *Anesthesiology*, v. 113 No 2 August 2010.

14. A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Composite postoperative morbidity at 30 days was also higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.

15. A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of women and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative

mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with a preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_02_evidence_attachment-635996215345848997.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
There are many corrective interventions available for patients identified with preoperative sub-optimal hemoglobin levels in order to avoid a transfusion during or after the surgical procedure. As an essential component of blood management, pre-operative investigation and correction of anemia should be undertaken, since transfusion has been shown to increase adverse outcomes. Early detection, evaluation, and management of preoperative anemia has been identified as an unmet medical need5.

One study of hip and knee arthroplasty patients found that those with a hemoglobin level <13.0g/dL. had four times the risk for blood transfusion than those with higher hemoglobin levels5.

Prevalence of preoperative anemia varies by population: Community-dwelling, >65 years old - <10%

i. Frail nursing home resident - >48%
ii. Surgical population – 5% to 75%
iii. Octogenarian, elective cardiac surgery – 49.4%1
iv. 7% of 9,462 patients undergoing total hip or total knee replacement2
v. >65 years old – 11% women, 10.2% men (NHANES Study)3
vi. Elective orthopedic surgery – 35%4

1. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
2. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
3. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
4. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.

Preoperative anemia is also a predictor of postoperative transfusion in orthopedic, major colon, and major cardiac surgery. Since blood transfusion is the most frequently-performed hospital procedure (11% of hospital stays) and has increased by 126% from 1997 – 2010, and since blood transfusion can have adverse outcomes, such as prolonged length of stay and decreased functional status at discharge, investigation and correction of preoperative anemia is essential to any blood management program.

The World Health Organization has defined the levels of anemia for men at a hemoglobin measurement of less than 13.0, and for non-pregnant women at a hemoglobin measurement of less than 12.0. There has, however, been controversy over these levels. While there is debate regarding the hemoglobin level at which patients are considered anemic7, use of the WHO definition of anemia allows identification of patients for whom pre-operative investigation and correction of hemoglobin levels is warranted.

The intent of the measure is to provide information to providers and review groups about the incidence of transfusions in the various strata, with the objective of identifying trends related to over- and underutilization of blood transfusions and correction of preoperative anemia.

5. Spahn DR. Anemia and Patient Blood Management in Hip and Knee Surgery. Anesthesiology, v. 113 No 2 August 2010.
6. Salido J et al. Preoperative Hemoglobin Levels and the Need for Transfusion After Prosthetic Hip and Knee Surgery. The Journal of Bone and Joint Surgery. Volume 84-A – Number2 – February 2002.
7. Beutler E, Waalen J. The definition of anemia: what is the lower limit of normal of the blood hemoglobin concentration? Blood Mar 1 2006 (107)5: 1747-1750.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This is a new measure for which approval for trial use is requested.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
Incidence of preoperative anemia –
b. Incidence of anemia increases with age but varies by subpopulation.
i. Community-dwelling, >65 years old - <10%
ii. Frail nursing home resident - >48%
iii. Surgical population – 5% to 75%
iv. Octogenarian, elective cardiac surgery – 49.4%1
v. 7% of 9,462 patients undergoing total hip or total knee replacement2
vi. >65 years old – 11% women, 10.2% men (NHANES Study)3
vii. Elective orthopedic surgery – 35%4

8. Partridge J, Harari D, Gossage J, Dhesi J. Anaemia in the older surgical patient: a review of prevalence, causes, implications and management. J R SOC Med 2013: 106: 269-277. (Literature review).
9. Bierbaum B et al. An Analysis of Blood Management in Patients Having a Total Hip or Knee Arthroplasty. The Journal of Bone and Joint Surgery Vol 81-A January, 1989 pp. 1-10.
10. Gurainek J et al. Prevalence of anemia in persons 65 years and older in the United States: evidence for a high rate of unexplained anemia. Blood. 2004;104: 2263 – 2268).
11. Goodnough, et al. Detection. Evaluation, and Management of Anemia in the Elective Surgical Patient. Anesth Analg 2005; 1858 – 61.
A retrospective cohort study of 227,425 patients undergoing major non-cardiac surgery in 2008 from the American College of Surgeons' National Surgical Quality Improvement Program database showed that 30.44% had preoperative anemia and after adjustment, postoperative mortality at 30 days was higher in patients with anemia than in those without anemia. Composite postoperative morbidity at 30 days was also higher in patients with anemia than in those without anemia. Musallam KM, Tamim HM, Richards T, Spahn DR, et al. Preoperative anemia and postoperative outcomes in non-cardiac surgery: a retrospective cohort study. Lancet 2001 Oct 15; 378(9800) 1396 – 407.
A single-center retrospective cohort study was conducted on 7,759 consecutive non-cardiac surgical patients between 2003 and 2008. 39.5% of women and 39.9% of women had preoperative anemia, and preoperative anemia was associated with a nearly five-fold increase in the odds of postoperative mortality. Beattie WS, Karkouti K, Wijaysundera DN, Tait G. Risk associated with a preoperative anemia in noncardiac surgery: a single-center cohort study. Anesthesiology. 2009 Mar;110(3): 574-81.
In addition, in a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 118 of the 141 respondents (81%) indicated that there was a gap in practice; 6 were not sure, and 17 reported no gap. Of the 118, most indicated that pre-operative anemia screening was done 3 or 4 days in advance of the elective surgical procedure. Given that 3-4 days is an insufficient period of time to correct any anemia, a high incidence of patients undergoing elective surgery with uncorrected anemia is presumed. Unpublished data.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*

No disparities are identified.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparities are identified in the literature.

1c. High Priority (previously referred to as High Impact)
The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Blood transfusion is the most common procedure performed during hospitalization,1 but research shows 50 percent of red blood cell transfusions are found to be inappropriate.2

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Most Frequent Procedures Performed in U.S. Hospitals, 2010, Healthcare Cost and Utilization Project (HCUP). February 2013. Agency for Healthcare Research and Quality.
2. Shander et al. Appropriateness of Allogeneic Red Blood Cell Transfusion: The International Consensus Conference on Transfusion Outcomes. Transfusion Medicine Reviews, Vol 25, No 3 (July), 2011: pp 232-246.e53.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not a PRO-PM

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.jointcommission.org/measure_development_initiatives.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-02_PreopHemoglobinLevel.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PreopHemoglobinLevel_v4_3_Wed_Jun_08_15.16.14_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients whose hemoglobin level measured on the most recent pre-operative hemoglobin level was:
12.0 grams or above
>=11.0 and <12.0 grams (mild anemia)
>=8.0 and <11.0 grams (moderate anemia)
Below 8.0 grams (severe anemia)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Pre-operative hemoglobin level is represented as a code from the following value set and associated QDM datatype:
"Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Selected elective surgical patients age 18 and over, who received a transfusion of whole blood or packed cells in the time window from anytime during the surgical procedure to 5 days after the surgical procedure or to discharge, whichever is sooner.

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Inpatient encounters are represented by the valueset and associated QDM datatype:
"Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
Selected elective surgical procedures are represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"
Transfusion of whole blood or packed cells is represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who are pregnant during the hospitalization, including those who delivered and those who did not deliver during this hospitalization
- Patients who undergo extra-corporeal membrane oxygenation procedures (ECMO) prior to the elective surgical procedure.
- Patients with sickle cell disease or hereditary hemoglobinopathy

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

Traumatic injury is represented by a code from the following value set and associated QDM datatype:

Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Solid organ transplant is represented by a code from the following value set and associated QDM datatype;
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"

Pregnancy, delivered and not delivered, is represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: Maternal and Fetal Procedures" using "Maternal and Fetal Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.51)

Or

Attribute: "Diagnosis: Pregnancy, Childbirth, and the Puerperium Grouping Value Set (2.16.840.1.113762.1.4.1029.50)

ECMO is represented by a code from the following value set and associated QDM datatype:
"Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)"

Sickle cell disease and hereditary hemoglobinopathy is represented by a code from the following value set and associated QDM datatype:
Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*

Stratification 1 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result >= 12.0 g)"

Stratification 2 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 11.0 g)
(result < 12.0 g)

Stratification 3 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 8.0 g)
(result < 11.0 g)

Stratification 4 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Selected Elective Surgical Procedures"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result < 8.0 g)"

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Count
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Score within a defined interval

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not a PRO-PM or survey

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM_02_testing_form_for_trial_use.docx,PBM02_CMS601v0_Bonnie_Export.xlsx

National Quality Forum

Measure Testing Form for Trial Approval Program

Measure Title: PBM-02: Preoperative Hemoglobin Level
Date of Submission: 5/31/2016
Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submission that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide preliminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the measure meets reliability and validity must be in this form

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*including questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

1. DATA/SAMPLE USED FOR PRELMINARY TESTING OF THIS MEASURE

It is important that the measure developer use a data set to conduct preliminary testing in order to evaluate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

The Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.

Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other:

1.4. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

78 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Bonnie testing

was also performed for each stratum specified in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions, as well as complex cases containing multiple values for hemoglobin results as well as multiple transfusions, to confirm the stratification logic performed as expected.

If the Bonnie testing tool was used to provide a sample data set, please refer to the guidance for Bonnie testing found at this link:
http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=80307 **Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. Refer to this link for an example of formatting Bonnie results:**
http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=81576

Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.

RELIABILITY AND VALIDITY ASSESSMENTS

Note: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) **Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program.**

Include descriptions of:

Inter-abstractor reliability, and data element reliability of all critical data elements

Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-02.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.38
Denominator clearly describes the activity being measured	4.46
Numerator inclusions clear and appropriate	4.51
Denominator inclusions clear and appropriate	4.53
Numerator exclusions clear and appropriate	4.44
Denominator exclusions clear and appropriate	4.45
Accurately assesses the process of care to which it is addressed	4.13

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Denominator test cases positively test to ensure that only patients who have a blood transfusion administered <=5 day(s) after the start of selected surgical procedures are included in the denominator. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patients who receive transfusion one minute after surgery, at exactly 5 days, and 6 days after surgery. Patients receiving transfusions one minute and five days after surgery were included in the denominator, while patients receiving transfusions at 6 days after surgery were not.

Numerator test cases positively test to ensure patients who have a hemoglobin result recorded <= 45 days(s) prior to the start of surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in which hemoglobin

results were recorded >45 days prior to surgery or after surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures or encounter diagnoses. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-01 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Maternal and Fetal procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- ECMO procedures recorded in SNOMEDCT or ICD10PCS that start prior to the elective surgical procedure
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing any of the following conditions:
    - Traumatic Injury
    - Pregnancy, Childbirth, and the Puerperium
    - Sickle Cell Disease and Related Blood disorders.

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**
Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment **Attachment:** PBM02_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
n/a

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Usability and Use Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.
Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)
There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public

domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

n/a

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new measure.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance

results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
This is a new measure for which approval for trial use is requested.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
n/a

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630---

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of public comment and testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List

Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health

Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care
Medicine, Division of Vascular, Thoracic, Transplant Anesthesia
Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.
Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center
Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America
Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic
Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic

Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center
Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 05, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:** Ad.6. Copyright Statement

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-02: Preoperative Hemoglobin Level

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements that were deemed to be more difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"
3. "Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
4. "Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)"
5. "Procedure, Performed: Maternal and Fetal Procedures" using "Maternal and Fetal Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.51)"
6. "Procedure, Performed: Selected Elective Surgical Procedures" using "Selected Elective Surgical Procedures Grouping Value Set (2.16.840.1.113762.1.4.1029.19)"

7. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
8. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
9. Attribute: "Diagnosis: Pregnancy, Childbirth, and the Puerperium" using "Pregnancy, Childbirth, and the Puerperium Grouping Value Set (2.16.840.1.113762.1.4.1029.50)"
10. Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

Initial Population and Denominator Data Elements

Data elements 1- “Encounter, Performed: Encounter Inpatient,” 3- “Procedure, Performed: Blood Transfusion Administration” and 6- “Procedure, Performed: Selected Elective Surgical Procedures” are used to define the initial population and denominator of this measure.

On the feasibility scorecard, hospitals rated these data elements 1 and 6 as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year timeframe outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Four out of five hospitals rated capture of data element 6 as feasible or highly feasible, represented as a score of 2 or 3 out of 3. Facilities rating the data element as a 2 cited variation in clinical workflow and adoption of new technology as reasons for the lower rating. One site rated current state feasibility as a 1, as it did not currently have an interface between the OR scheduling system where this information was captured and the certified EHR technology. This site had plans to transition to an interfaced OR module in 1-2 years. All site rated the future state as highly feasible.

Finally, four out of five sites rated data element 3 as highly feasible, represented as a score of 3 out of 3. One hospital rated current feasibility as a 1 as blood transfusion was documented on a paper record, but had plans to implement blood transfusion via the EHR within six months.

Numerator Data Element

Data element 2- “Laboratory Test, Performed: Hemoglobin blood serum plasma” is used to define the numerator for this measure. Specifically, cases with a hemoglobin result recorded within 45 days prior to the start time of an elective surgical procedure meet numerator criteria. This time frame differs slightly from PBM-01, which evaluates hemoglobin results captured 14-45 days prior to surgery. In this measure, hemoglobin results up to the day of surgery meet numerator criteria.

Hospitals reported that hemoglobin results are routinely captured as structured data prior to surgery. However, limited interoperability between hospitals and their community partners, such as clinics and lab centers, limits the availability of structured data for lab results from external laboratories. Hospitals noted that many external results are received via fax, or as an electronic document, rather than in a format that can be structured and encoded in the EHR.

Hospitals rated feasibility of capturing this data element as high in most circumstances, but noted the interoperability issues may currently impact the availability of data for some cases.

Denominator Exclusions Data Elements

Data elements 4, 5, 7, 8, 9, and 10 are used to represent denominator exclusions.

Feasibility for data elements 4-“Procedure, Performed: ECMO,” 5-“Procedure Performed, Maternal and Fetal Procedures,” and 7-“Procedure, Performed: Solid Organ Transplant,” was found to be comparable to 3-“Procedure, Performed: Selected Elective Surgical Procedures.” These data elements are found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform solid organ transplant, which would not use this data element.

Data elements 8- “Attribute, Diagnosis: Traumatic Injury,” 9- “Attribute: Diagnosis: Pregnancy, Childbirth, and the Puerperium,” and 10-“Attribute: “Diagnosis: Sickle Cell Disease and Related Blood Disorders” represent encounter diagnoses. All hospitals rated these data elements as highly feasible. Discussion around these data elements suggested that the functionality to support collection of these data elements are well established.

Conclusion

Hospitals completing the feasibility scorecard largely reported the data elements required to calculate this measure to be feasible or highly feasible in the current state, with the exception of the numerator data element representing hemoglobin results. Capture of hemoglobin results from external laboratories will require improvements in interoperability or workarounds to support data collection. Approval for Trial Use status will support The Joint Commission’s efforts to further test this measure.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3019
De.2. **Measure Title:** PBM-03: Preoperative Blood Type Testing and Antibody Screening
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure assesses the proportion of selected elective surgical patients age 18 and over who had timely preoperative assessment of blood type and crossmatch or type and screening.
1b.1. **Developer Rationale:** Beginning an elective surgery that is considered a high-blood loss procedure without confirming the availability of a patient's specific blood type should be an important patient safety concern for all hospitals and patients. Hospitals need to ensure that sufficient compatible blood is available for each scheduled procedure since about 3% of specimens have a serologic finding that requires further investigation that may cause a delay in the availability of the blood. Now that many patients do not have blood testing until the day of the procedure, the results may not be completed by the time surgery begins. Studies related to the timely completion of type and screen (T&S) and verification of ABO/Rh status for elective surgery patients were minimal, but one recent study showed that 21 (7%) of the 309 patients scheduled for elective surgery, did not have the T&S sample tested before surgery.1 In another study, type and screen collected less than 3 days prior to surgery resulted in special efforts to find blood more than 1% of the time; type and screen collected on the same day as surgery resulted in a surgery delay almost 1% of the time.2

According to the 2011 Joint Commission's National Patient Safety Goal UP.01.01.01., a pre-procedure verification process should be conducted to identify items that must be available for the procedure using a standardized list that includes documentation of any required blood products for the procedure. Development of formal protocols to ensure that patients have blood testing completed (when ordered) prior to anesthesia start time for potential high-blood loss elective surgeries may optimize management of blood resources and maximize patient safety. The benefits envisioned by the use of this measure are a reduction in transfusion of uncross=matched blood, a reduction in transfusion reactions, and reduced surgical delays.

1. Chiganti S, Regan F. Are changes in admission practices for elective surgery posing a transfusion threat to patients? Transfusion Medicine, 2002, 12, 353-356.
2. Friedberg KC, Jones BA, Walsh MK. Type and Screen Completion for Scheduled Surgical Procedures. Arch Pathol Lab Med – Vol 127, May 2003.

S.4. **Numerator Statement:** Patients who had a type and crossmatch or type and screen completed within 45 days prior to the surgery start date and time.
S.7. **Denominator Statement:** Selected elective surgical patients age 18 and over
S.10. **Denominator Exclusions:** • Patients under age 18
• Patients whose surgical procedure is performed to address a traumatic injury
• Patients who have a solid organ transplant
• Patients who refuse transfusion

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☐ **Yes** ☒ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☐ **Yes** ☒ **No**
- **Evidence graded?** ☐ **Yes** ☒ **No**

Evidence Summary

The developer provides the following path to support the relationship between the process of care (timely preoperative assessment of blood type and crossmatch or type and screening) and outcomes:

1. Process: Timely performance of type and crossmatch or type and screen
2. Availability of cross-matched blood when needed
3. Outcomes: 1. Avoidance of surgery delays 2. Avoidance of hemolytic or other adverse reactions due to transfusion of incompatible blood.

Two clinical guidelines sources are cited to support the measure; the Red Cross guideline recommendation states the following: *RBCs must be compatible with ABO antibodies present in the recipient plasma and must be crossmatched (serologically or electronically, as applicable) to confirm compatibility with ABO and other clinically significant antibodies prior to routine transfusion. Units must be negative for the corresponding antigens.* Additionally, the American Society of Hematology's recommendation is: *Pretransfusion Testing: "Age of sample – Longer (than 3 days, often 1-2 weeks depending on hospital policy) for outpatient pre-op testing if negative history (for pregnancy or transfusion recipient) within 3 months.* The guideline recommendations are not graded.

The developer did not provide specific evidence to support the 45 day prior to surgery timeframe for a type and crossmatch or a type and screen.

Exception to evidence

NQF guidance provides that when there is insufficient empirical evidence in support of a measure, a determination should be made about whether there are, or could be, performance measures of a related outcome or evidence-based intermediate clinical outcome or process. Currently, there are no NQF-endorsed measures related to pre-operative blood type testing and antibody screening. An example of such a measure might be the proportion of selected elective surgical patients ages 18 and over who had timely preoperative assessment of blood type and crossmatch or type and screening who received blood transfusion post-surgery. If the Committee accepts that there is an appropriate alternate measure (the example or some variant), an exception to the evidence would not be warranted. Alternatively, if the Committee does not identify an appropriate alternate measure, it may agree that it is OK (beneficial) to hold providers accountable for performance in the absence of empirical evidence of benefits to patients, in which case it would rate the evidence as insufficient with exception.

Guidance from the Evidence Algorithm

Process measure supported by guideline recommendation; unclear if based on systematic review (Box 3) → Evidence not graded (Box 7) → A measure of a related outcome may not exist (Box 10) → Systematic assessment of expert opinion (Box 11) → If Committee agrees it is OK/beneficial to hold providers accountable for performance in the absence of empirical evidence of benefits to patients → rate as INSUFFICIENT WITH EXCEPTION

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

○ *For possible exception to the evidence criterion:*

- *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
- *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*
- *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Although there is no performance data on the measure as specified, the developer provided a summary of data from four difference sources that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.

Disparities

- The developed indicated that no disparity data are available.

Questions for the Committee:

○ *Is there a gap in care that warrants a national performance measure?*

○ *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

- Process Measure
  . Numerator Statement: Patients who had a type and crossmatch or type and screen completed within 45 days prior to the surgery start date and time.
  Denominator Statement: Selected elective surgical patients age 18 and over
  PRELIMINARY RATING : INSUFFICIENT

- This is a process measure. what is the outcome of concern? Should that be measured instead?

Again, concern with which procedures should have a T&S done

1b. Performance Gap

- PERFORMANCE GAP: MODERATE one recent study showed that 21 (7%) of the 309 patients scheduled for elective surgery, did not have the T&S sample tested before surgery.1 In another study, type and screen collected less than 3 days prior to surgery resulted in special efforts to find blood more than 1% of the time; type and screen collected on the same day as surgery resulted in a surgery delay almost 1% of the time

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients who had a type and crossmatch or type and screen completed within 45 days prior to the surgery start date and time.
- Denominator Statement: Selected elective surgical patients age 18 and over
- Denominator Exclusions:
  - Patients under age 18
  - Patients whose surgical procedure is performed to address a traumatic injury
  - Patients who have a solid organ transplant
  - Patients who refuse transfusion
- Level of Analysis: Facility
- Care Setting: Hospital/Acute Care Facility
- No risk adjustment or risk stratification

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously; Submitted Bonnie test results
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high

proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 25 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications are consistent with evidence? Does the evidence support the 45-day prior to surgery timeframe?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.41
Denominator clearly describes the activity being measured	4.39
Numerator inclusions clear and appropriate	4.39
Denominator inclusions clear and appropriate	4.04
Numerator exclusions clear and appropriate	4.41
Denominator exclusions clear and appropriate	4.36
Accurately assesses the process of care to which it is addressed	4.22

This measure is being considered for trial use, thus full validity testing results are not expected and the Committee will not vote on this criterion.

2b3-2b7. Threats to Validity

2b3. Exclusions:

When data are available, the developer plans to analyze exclusion frequency and variability across providers in order to determine if excluded cases affect overall performance scores for the measure. The data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury
- Patients who refuse transfusion

Questions for the Committee:

- *Are there other threats to validity the measure developer should consider?*
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time.

2b6. Comparability of data sources/methods:

N/A

2b7. Missing Data

The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications: Reliability-Specifications

- Clear descriptors However maybe hard to have this measure consistently implemented without incentives

  Outcomes: 1. Avoidance of surgery delays 2. Avoidance of hemolytic or other adverse reactions due to transfusion of incompatible blood

2a2. Reliability – Testing

- The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use;

2b.1 Validity – Specifications

- Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
  S.26. Level of Analysis: Facility

2b2. Validity – Testing

- VALIDITY : SOMEWHAT Non specific guidelines : the Red Cross guideline recommendation states the following: RBCs must be compatible with ABO antibodies present in the recipient plasma and must be crossmatched (serologically or electronically, as applicable) to confirm compatibility with ABO and other clinically significant

antibodies prior to routine transfusion. Units must be negative for the corresponding antigens. Additionally, the American Society of Hematology's recommendation is: Pretransfusion Testing: "Age of sample – Longer (than 3 days, often 1-2 weeks depending on hospital policy) for outpatient pre-op testing if negative history (for pregnancy or transfusion recipient) within 3 months. The guideline recommendations are not graded.

2b3-7. Threats to Validity

- The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- MODERATE • The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**

OR

Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms N/A

Feedback :

None identified

***Questions for the Committee**:*

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- CURRENTLY NOT PUBLICLY REPORTED AND NOT IN AN ACCOUNTABLE PROGRAM BUT YES TO POSSIBLY IN FUTURE Outcomes: 1. Avoidance of surgery delays 2. Avoidance of hemolytic or other adverse reactions due to transfusion of incompatible blood

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

-

Submission materials attachments…

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: PBM-03: Preoperative Type and Crossmatch or Type and Screen

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Timely performance of of type and crossmatch or type and screen prior to elective surgery for inpatients age 18 and over. .

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

5. Process: Timely performance of type and crossmatch or type and screen
6. Availability of cross-matched blood when needed
7. Outcomes: 1. Avoidance of surgery delays 2. Avoidance of hemolytic or other adverse reactions due to transfusion of incompatible blood.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

Guideline #1.

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Vassallo R, et al. A Compendium of Transfusion Practice Guidelines, Second Edition, 2013. American Red Cross, page 8.
http://www.redcrossblood.org/sites/arc/files/59802_compendium_brochure_v_6_10_9_13.pdf

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page 8: RBCs must be compatible with ABO antibodies present in the recipient plasma and must be crossmatched (serologically or electronically, as applicable) to confirm compatibility with ABO and other clinically significant antibodies prior to routine transfusion. Units must be negative for the corresponding antigens.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade not assigned.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

n/a

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

n/a

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

Guideline #2.

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Weinstein R. 2012 Clinical Practice Guideline on Red Blood Cell Transfusion. Presented by the American Society of Hematology, adapted from "Red Blood Cell Transfusion: A Clinical Practice Guideline from the AABB". *Ann Intern Med*. 2012; 157: 49-58.

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Guideline #3: Pretransfusion Testing: "Age of sample – Longer (than 3 days, often 1-2 weeks depending on hospital policy) for outpatient pre-op testing if negative history (for pregnancy or transfusion recipient) within 3 months."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Grade not assigned.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

n/a

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

n/a

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, MEDLINE and other relevant sources including professional association websites, The Cochrane Library, the National Guideline Clearinghouse, and other sources was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Dean L. Blood Groups and Red Cell Antigens. 2005. National Center for Biotechnology Information, U.S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, 20894 USA
    a. If incompatible blood is given in a transfusion, the donor cells are treated as if they were foreign invaders, and the patient's immune system attacks them accordingly. Not only is the blood transfusion rendered useless, but a potentially massive activation of the immune system and clotting system can cause shock, kidney failure, circulatory collapse, and death.
    b. To avoid a transfusion reaction, donated blood must be compatible with the blood of the patient who is receiving the transfusion. More specifically, the donated RBCs must lack the same ABO and Rh D antigens that the patient's RBCs lack.

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 subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_01_evidence_attachment.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Beginning an elective surgery that is considered a high-blood loss procedure without confirming the availability of a patient's specific blood type should be an important patient safety concern for all hospitals and patients. Hospitals need to ensure that sufficient compatible blood is available for each scheduled procedure since about 3% of specimens have a serologic finding that requires further investigation that may cause a delay in the availability of the blood. Now that many patients do not have blood testing until the day of the procedure, the results may not be completed by the time surgery begins. Studies related to the timely completion of type and screen (T&S) and verification of ABO/Rh status for elective surgery patients were minimal, but one recent study showed that 21 (7%) of the 309 patients scheduled for elective surgery, did not have the T&S sample tested before surgery.1 In another study, type and screen collected less than 3 days prior to surgery resulted in special efforts to find blood more than 1% of the time; type and screen collected on the same day as surgery resulted in a surgery delay almost 1% of the time.2

According to the 2011 Joint Commission's National Patient Safety Goal UP.01.01.01., a pre-procedure verification process should be conducted to identify items that must be available for the procedure using a standardized list that includes documentation of any required blood products for the procedure. Development of formal protocols to ensure that patients have blood testing completed (when ordered) prior to anesthesia start time for potential high-blood loss elective surgeries may optimize management of blood resources and maximize patient safety. The benefits envisioned by the use of this measure are a reduction in transfusion of uncross=matched blood, a reduction in transfusion reactions, and reduced surgical delays.

1. Chiganti S, Regan F. Are changes in admission practices for elective surgery posing a transfusion threat to patients? Transfusion Medicine, 2002, 12, 353-356.
2. Friedberg KC, Jones BA, Walsh MK. Type and Screen Completion for Scheduled Surgical Procedures. Arch Pathol Lab Med – Vol 127, May 2003.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This is a new measure for which approval for trial use is being requested.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
A. In a study of 309 patients undergoing elective major surgical procedures in one institution, of the 187 patients for whom a Type and Screen was indicated, 20 (10.7%) had no results available by the start of surgery. Chiganti S, Regan F. Are changes in admission practices for elective surgery posing a transfusion threat to patients? Transfusion Medicine, 2002, 12, 353-356.

B. A Q-Probe study by the College of American Pathologists (CAP) reviewed 108 public and private participating institutions for performance, completion, and results of Type and Screen testing of scheduled elective surgical patients. Of the 8941 type and screens, 64.6% were collected prior to the day of surgery. Of those type and screens collected the day of surgery the median laboratory completed almost 23% after the start of surgery. For 10% of participants, more than 73% of all type and screens collected on the same day as surgery were not complete until after the start of surgery. When red blood cell-directed antibodies were identified, 78.7% were considered clinically significant. Positive antibody screens were significantly associated with delayed surgery and special efforts to obtain blood. Type and screen collected on the same day as surgery directly resulted in a surgery delay 0.8% of the time. Friedberg KC, Jones BA, Walsh MK. Type and Screen Completion for Scheduled Surgical Procedures. Arch Pathol Lab Med – Vol 127, May 2003.

C. During a 4-month period, serological problems arose in 70 of 2859 cases. In 36 of the 70 cases, the sample arrived at the blood bank about the time of the beginning of the operation; in 19 of these 36 cases, the operation had begun before serological resolution and in 7 of these 19, the antibody was found to be of hemolytic potential. Moore SB, Reisner RK, Losasso TJ, Brockman SK. Morning admission to the hospital for surgery the same day. A practical problem for the blood bank. Transfusion 1987 Jul-Aug; 27 (4) 359-61.

D. In a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 94 of the 141 respondents (66.7%) indicated that there was a gap in practice; 31 were not sure, and 16 reported no gap.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
No disparities data are identified.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparity data are identified in the literature

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Patient/societal consequences of poor quality
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
a. Blood transfusion is the most common procedure performed during hospitalization,1

b. In a study of 309 patients undergoing elective major surgical procedures in one institution, of the 187 patients for whom a Type and Screen was indicated, 20 (10.7%) had no results available by the start of surgery.2
c. A Q-Probe study by the College of American Pathologists (CAP) reviewed 108 public and private participating institutions for performance, completion, and results of Type and Screen testing of scheduled elective surgical patients. Of the 8941 type and screens, 64.6% were collected prior to the day of surgery. Of those type and screens collected the day of surgery the median laboratory completed almost 23% after the start of surgery. For 10% of participants, more than 73% of all type and screens collected on the same day as surgery were not complete until after the start of surgery. When red blood cell-directed antibodies were identified, 78.7% were considered clinically significant. Positive antibody screens were significantly associated with delayed surgery and special efforts to obtain blood. Type and screen collected on the same day as surgery directly resulted in a surgery delay 0.8% of the time.3
d. During a 4-month period, serological problems arose in 70 of 2859 cases. In 36 of the 70 cases, the sample arrived at the blood bank about the time of the beginning of the operation; in 19 of these 36 cases, the operation had begun before serological resolution and in 7 of these 19, the antibody was found to be of hemolytic potential.4
e. In a survey of 141 acute-care hospitals conducted by The Joint Commission in 2015, respondents were asked if there was a gap between their current practice and the parameters proposed by this measure. 94 of the 141 respondents (66.7%) indicated that there was a gap in practice; 31 were not sure, and 16 reported no gap. Unpublished data.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Most Frequent Procedures Performed in U.S. Hospitals, 2010, Healthcare Cost and Utilization Project (HCUP). February 2013. Agency for Healthcare Research and Quality.
2. Chiganti S, Regan F. Are changes in admission practices for elective surgery posing a transfusion threat to patients?

Transfusion Medicine, 2002, 12, 353-356.
3. Friedberg KC, Jones BA, Walsh MK. Type and Screen Completion for Scheduled Surgical Procedures. Arch Pathol Lab Med – Vol 127, May 2003.
4. Moore SB, Reisner RK, Losasso TJ, Brockman SK. Morning admission to the hospital for surgery the same day. A practical problem for the blood bank. Transfusion 1987 Jul-Aug; 27 (4) 359-61.
5. Unpublished data

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not a PRO-PM

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.jointcommission.org/measure_development_initiatives.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-03_PreoperativeBloodTypeTesting.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PreoperativeBloodTypeTesting_v4_3_Wed_May_25_08.46.30_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had a type and crossmatch or type and screen completed within 45 days prior to the surgery start date and time.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had a type and crossmatch or type and screen are represented by code in the following value set and associated QDM datatype:

- Laboratory Test, Performed: Blood Group Antibody Screen" using "Blood Group Antibody Screen LOINC Value Set (2.16.840.1.113762.1.4.1029.30)"
- "Laboratory Test, Performed: Major Crossmatch" using "Major Crossmatch LOINC Value Set (2.16.840.1.113762.1.4.1029.29)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Selected elective surgical patients age 18 and over

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Selected elective surgical patients are represented by a code in the following value set and associated QDM datatype:

"Procedure, Performed: Selected Elective Surgical Procedures PBM03" using "Selected Elective Surgical Procedures PBM03 Grouping Value Set (2.16.840.1.113762.1.4.1029.14)"

Inpatients age 18 and over are represented by a code from the following Value Set and associated QDM Datatype: "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who refuse transfusion

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Traumatic injury is represented by a code in the following value set and associated QDM datatype:

Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Solid organ transplant is represented by a code from the following value set and asscoiated QDM datatype:
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"

Refusal of transfusion is represented by a code from the following values set and associated QDM datatype:

"Procedure, Order not done: Patient Refusal" using "Patient Refusal SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.93)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
This measure is not stratified.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
n/a

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Se attached HQMF file.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Measure is not based on a survey; not a PRO-PM.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.

Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM_03_testing_form_for_trial_use.docx,PBM03_CMS606V0_Bonnie_Export.xlsx

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.jointcommission.org/measure_development_initiatives.aspx

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-03_PreoperativeBloodTypeTesting.zip

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** PreoperativeBloodTypeTesting_v4_3_Wed_May_25_08.46.30_CDT_2016.xls

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had a type and crossmatch or type and screen completed within 45 days prior to the surgery start date and time.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients who had a type and crossmatch or type and screen are represented by code in the following value set and associated QDM datatype:

- Laboratory Test, Performed: Blood Group Antibody Screen" using "Blood Group Antibody Screen LOINC Value Set (2.16.840.1.113762.1.4.1029.30)"
- "Laboratory Test, Performed: Major Crossmatch" using "Major Crossmatch LOINC Value Set (2.16.840.1.113762.1.4.1029.29)"

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
Selected elective surgical patients age 18 and over

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Selected elective surgical patients are represented by a code in the following value set and associated QDM datatype:

"Procedure, Performed: Selected Elective Surgical Procedures PBM03" using "Selected Elective Surgical Procedures PBM03 Grouping Value Set (2.16.840.1.113762.1.4.1029.14)"

Inpatients age 18 and over are represented by a code from the following Value Set and associated QDM Datatype: "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

- Patients under age 18
- Patients whose surgical procedure is performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients who refuse transfusion

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Traumatic injury is represented by a code in the following value set and associated QDM datatype:

Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Solid organ transplant is represented by a code from the following value set and asscoiated QDM datatype:
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"

Refusal of transfusion is represented by a code from the following values set and associated QDM datatype:

"Procedure, Order not done: Patient Refusal" using "Patient Refusal SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.93)"

S.12. Stratification Details/Variables *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
This measure is not stratified.

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
n/a

S.16. Type of score:
Rate/proportion
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Se attached HQMF file.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. Survey/Patient-reported data *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Measure is not based on a survey; not a PRO-PM.

S.22. Missing data (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the

structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM_03_testing_form_for_trial_use.docx,PBM03_CMS606V0_Bonnie_Export.xlsx

National Quality Forum

Measure Testing Form for Trial Approval Program

Measure Title: Preoperative Blood Type Testing and Antibody Screening

Date of Submission: 5/31/2016

Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submission that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide preliminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the measure meets the reliability and validity must be in this form.

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*including questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

1. DATA/SAMPLE USED FOR PRELMINARY TESTING OF THIS MEASURE

It is important that the measure developer use a data set to conduct preliminary testing in order to evaluate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can use can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

The BONNIE testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.

Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.,*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other:

1.4. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

25 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions.

All 25 cases passed or failed as expected based on the data included in the case, confirming the measure logic is accurate and valid. For further information on the characteristics of the patients included in the analysis, please refer to the attached BONNIE testing spreadsheet.

1.5. Please refer to the guidance for Bonnie testing found at this link. Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. *Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.*

Please refer to the attached BONNIE testing spreadsheet.

<u>RELIABILITY AND VALIDITY ASSESSMENTS</u>

<u>Note</u>: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program. Include descriptions of:

- Inter-abstractor reliability, and data element reliability of all critical data elements
- Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-03.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters for PBM-03: Preoperative Blood Type Testing and Antibody Screening:

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.41
Denominator clearly describes the activity being measured	4.39
Numerator inclusions clear and appropriate	4.39
Denominator inclusions clear and appropriate	4.04
Numerator exclusions clear and appropriate	4.41
Denominator exclusions clear and appropriate	4.36
Accurately assesses the process of care to which it is addressed	4.22

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Initial Population and Denominator test cases positively test to ensure that only patients >= 18 years of age who have a surgical procedure performed <=48 hours prior to the inpatient encounter or during the inpatient encounter are included. Negative test cases ensure that patients who do not meet these criteria do not pass into the denominator. For example, cases test patients who have a surgical procedure at 49 hours and 48 hours prior to the start of the encounter. Patients who have a surgical procedure 48 hours prior to the start of the encounter were included in the denominator, while patients with a surgical procedure at 49 hours prior to the encounter were not.

Numerator test cases positively test to ensure patients who have a major crossmatch or blood group antibody screen recorded <= 45 days prior to the start of surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases

in blood group antibody screens were recorded >45 days prior to surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures, encounter diagnoses, or refusal of blood transfusion. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-03 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers in order to determine if excluded cases affect overall performance scores for the measure. The data elements to be analyzed include:

- Solid Organ Transplant procedures recorded in SNOMEDCT or ICD10PCS that occur <=48 hours prior to admission or during the inpatient encounter.
- Encounter diagnoses recorded in SNOMEDCT or ICD10CM representing traumatic injury
- Patients who refuse transfusion

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PBM03_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

n/a

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)

There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

This is a new measure.

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new measure.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
This is a new measure for which approval for trial use is being requested.

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
n/a

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

Attachment:

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List

Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health

Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care
Medicine, Division of Vascular, Thoracic, Transplant Anesthesia
Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.
Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center
Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America
Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic
Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic

Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center
Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 05, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:** Ad.6. Copyright Statement

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-03: Preoperative Blood Type Testing and Antibody Screening

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements which test sites deemed to be difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Blood Group Antibody Screen" using "Blood Group Antibody Screen LOINC Value Set (2.16.840.1.113762.1.4.1029.30)"
3. "Laboratory Test, Performed: Major Crossmatch" using "Major Crossmatch LOINC Value Set (2.16.840.1.113762.1.4.1029.29)"
4. "Procedure, Order: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
5. "Procedure, Order not done: Patient Refusal" using "Patient Refusal SNOMEDCT Value Set (2.16.840.1.113883.3.117.1.7.1.93)"
6. "Procedure, Performed: Selected Elective Surgical Procedures PBM03" using "Selected Elective Surgical Procedures PBM03 Grouping Value Set (2.16.840.1.113762.1.4.1029.14)"

7. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
8. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"

Initial Population and Denominator Data Elements

Data elements 1- "Encounter, Performed: Encounter Inpatient," and 6- ""Procedure, Performed: Selected Elective Surgical Procedures PBM03" are used to define the initial population and denominator of this measure. On the feasibility scorecard, hospitals rated these data elements as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year timeframe outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Four out of five hospitals rated capture of data element 6 as feasible or highly feasible, represented as a score of 2 or 3 out of 3. Those rating feasibility as a 2 cited variations in practice and adoption of new technology as reasons for variations in practice and the data available. One site rated current state feasibility as a 1, as it did not currently have an interface between the OR scheduling system where this information was captured and the certified EHR technology. This site had plans to transition to an interfaced OR module in 1-2 years. All site rated the future state as highly feasible.

Numerator Data Elements

Data Elements 2- "Laboratory Test, Performed: Blood Group Antibody Screen," and 3- "Laboratory Test, Performed: Major Crossmatch," are used to define the numerator of this measure. Patients with either data element recorded will be included in the numerator.

Feasibility for data elements 2 and 3 were assessed separately with each site. However, given the data elements share a data source, results for the two data elements were identical. All five sites rated data availability, accuracy, definition, and standards as highly feasible, represented as a score of 3 out of 3. Two sites noted a practice gap at their facilities, and reported current workflow feasibility as a 2, noting that implementation of the measure would provide a mechanism for measuring and improving the gap in care.

Denominator Exclusion Data Elements

Patients with documentation of data elements 4 and 5, or 6, or 7, or 8 are excluded from the measure.

Data element 4 represents blood transfusion, and data element 5 represents patient refusal. Both must be present in a record for a patient to meet the exclusion for patient refusal of blood transfusion. Four out of five sites rated data element 4 as highly feasible, represented as a score of 3 out of 3. One hospital rated current feasibility as a 1 as blood transfusion was documented on a paper record, but had plans to implement blood transfusion via the EHR within six months. Organizations reported greater variability, and lower feasibility, for capturing patient refusal of blood transfusion as structured data. Limitation to feasibility primarily stemmed from paper-based blood transfusion consent process and the limited support EHRs provide in general for consent. However, all sites discussed options for adding this documentation to their EHR, and felt that capture of refusal would be feasible in the future state.

Please refer to Appendix A for further findings related to patient refusal of transfusion.

Feasibility for data element 7- "Procedure, Performed: Solid Organ Transplant" was found to be comparable to data element 6- "Procedure, Performed: Selected Elective Surgical Procedures PBM03." These data elements are both found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform organ transplant, which would not use this data element.

All hospitals rated data element 8- "Diagnosis: Traumatic Injury," as highly feasible. Discussion around this data element suggested that while missing data may occur due to physician practice related to updating the problem list, the functionality to support collection of this data element was well established.

Conclusion

Hospitals completing the feasibility scorecard largely reported the data elements required to calculate this measure to be feasible or highly feasible in the current state, with the exception of a denominator exclusion for patient refusal of transfusion, which was not currently implemented in the EHR but found to be feasible and worthwhile to add. Some sites were in the process of addressing interoperability issues linking operative records and transfusion records with the certified EHR, but had near term plans for completing this transition. In conclusion, The Joint Commission considers *PBM-03: Preoperative Blood Type Testing and Antibody Screening* to be a highly feasible measure specification.

Appendix A: Feasibility Scorecard Findings for Patient Refusal of Blood Transfusion

Site		Workflow		Data Availability		Data Accuracy		Data Element Definition		Data Standard	
		Score	Comments	Score	Comments	Score	Comments	Score	Comments	Score	Comments
1	Current	1	Two types of refusal- refused in the moment, or refused across the board	1	Overall consent on chart- notes only in EHR	1		1		1	
	Future (3-5 years)	2		3	Could capture as a precaution order in the EHR	3		3		3	
2	Current	2	Blood bank has an alert, entered by blood bank staff	1		2		1		1	
	Future (3-5 years)	3	Are discussing having a "Bloodless Medicine" order that could trigger an alert within each encounter.	3		3		3		3	
3	Current	1		1		2		1		1	
	Future (3-5 years)	3	Capture in ICD-10	3		3		3		3	
4	Current	2		1		3		3		1	
	Future (3-5 years)	2	Uncertain- depends on how it would be defined in the EHR. Ideally, would have electronic signature. Would question accuracy of a flowsheet row.	2	If we had electronic signature, would be accurate. (Is electronic signature a priority? Unknown)	3		3		3	
5	Current	2	Using FYI functionality in Epic because it crosses encounters. This is a category list and currently you cannot attach a code. Currently also captured on paper on consent as well as the initial agreement.	1		3		3		1	Not able to capture in EHR product

	Future (3-5 years)	3		2	Dependent on vendor development	3		3		2	Dependent on vendor developmen t

NATIONAL QUALITY FORUM

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3020
De.2. **Measure Title:** PBM-04: Initial Transfusion Threshold
Co.1.1. **Measure Steward:** The Joint Commission
De.3. **Brief Description of Measure:** This measure assesses the proportion of various pre-transfusion hemoglobin levels in patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion. Over time, in a patient blood management program, there should be a higher proportion of patients receiving blood at the lower hemoglobin threshold and a lower proportion receiving blood at the higher hemoglobin thresholds. It also identifies patients who receive transfusions that should be reviewed by hospital transfusion/blood usage committees so that appropriate educational programs can be developed as part of a patient blood management program.
1b.1. **Developer Rationale:** All published sources indicate that a strict transfusion strategy is preferable to a liberal strategy, since transfusion can be harmful and contributes to higher mortality, infection, and other complications.1,2,3,4 Most guidelines recommend a threshold of 7.0 or 8.0 grams of hemoglobin or less as an indication for transfusion, and if the hemoglobin level is 10.0 or greater there is agreement that the transfusion is rarely indicated. There is disagreement among guidelines, however, when patients have underlying cardiac disease, postoperative status, or other clinical conditions5, others, and there is some concern about a lack of robust evidence to support some guidelines.6,7 Caution is always advised that when determining the appropriateness of transfusion, underlying clinical conditions and symptoms should be taken into consideration.
The purpose of this measure is to allow facilities to profile blood usage according to initial transfusion hemoglobin thresholds. Strata are defined to direct facility review to the appropriateness of selected transfusions, taking into account clinical symptoms combined with hemoglobin measurements. By this review, facilities will be able to determine the best approaches to enhance blood conservation and management and over time, there should be a gradual decline in the proportion of initial units given from the higher hemoglobin values to those lower values supported in a restrictive transfusion strategy in the literature and guidelines as part of a blood management program.

1. Carson JL, Grossman BJ, Kleinman S, Tinmouth at, et al. Red Blood Cell Transfusion: A Clinical Practice Guideline From the AABB. Ann Intern Med. 2012;157(1):49-58.
2. Goodnough LT, Shander A. Patient Blood Management. Anesthesiology v116; No 6, June 2012
3. Paone G, Likosky DS, Brewer R, Theurer PF, et al. Transfusion of 1 and 2 Units of Red Blood Cells is Associated With Increased Morbidity and Mortality. Ann Thorac Surg 2014; 97:87-94.
4. Shander A, Goodnough LT. Can Blood Transfusion Be Not Only Ineffective, But Also Injurious? Ann Thorac Surg 2014; 97: 11-4.
5. Shander A, Gross I, Hill S, Javidroozi M, Sledge S. A new perspective on best transfusion practices. Blood Transfus 2013; 11: 193-202.
6. Carson JL, Carless PA, Hebert PC. Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion (review). The Cochrane Collaboration, April 2012.
7. Wilkinson KL, Brunskill SJ, Doree C, Hopewell S, et al.The Clinical Effects of Red Blood Cell Transfusions: An Overview of the Randomized Controlled Trials evidence Base. Transfusion Medicine Reviews, Vol 25, No.2 (April), 2011, pp 145-155 e2.

S.4. **Numerator Statement:** Patients whose hemoglobin level measured prior to the transfusion and closest to the transfusion was:
- less than 7.0 grams
- >=7.0 and <8.0 grams
- >=8.0 and <9.0 grams
- >=9.0 and <10.0 grams
- 10.0 grams or greater

S.7. **Denominator Statement:** Patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion
S.10. **Denominator Exclusions:** • Patients who have a surgical procedure performed to address a traumatic injury
• Patients who have a solid organ transplant
• Patients undergoing extracorporeal membrane oxygenation (ECMO) treatment at the time of initial transfusion.
• Patients whose first unit of whole blood or packed red blood cells was given while an Emergency Department patient.
• Patients with sickle cell disease or hereditary hemoglobinopathy

De.1. **Measure Type:** Process
S.23. **Data Source:** Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory
S.26. **Level of Analysis:** Facility

IF Endorsement Maintenance – Original Endorsement Date: Most Recent Endorsement Date:

IF this measure is included in a composite, NQF Composite#/title:

IF this measure is paired/grouped, NQF#/title:

De.4. **IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?** This measure is not paired or grouped.

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- The developer provides the following path to support assessment of the proportion of various pre-transfusion hemoglobin levels in patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion and outcomes:
  1. Process: Assessment of pre-transfusion hemoglobin level, and administration of transfusion if hemoglobin below 7 or 8, or if clinical presentation requires transfusion at higher hemoglobin levels (restrictive strategy).
  2. Outcomes: A. Avoidance of transfusion and attendant complications. B. Reduced length of stay, reduced morbidity and mortality. C. Blood resource conservation.
- The rationale for the measure is supported by clinical guideline recommendations:
  1. AABB:
     - *Recommendation 1: The AABB recommends adhering to a restrictive transfusion strategy (7 to 8 g/dL) in hospitalized, stable patients. (*Grade: strong recommendation; high-quality evidence.)
     - *Recommendation 2: The AABB suggests adhering to a restrictive strategy in hospitalized patients with preexisting cardiovascular disease and considering transfusion for patients with symptoms or a hemoglobin level of 8 g/dL or less. (*Grade: weak recommendation; moderate-quality evidence.)
     - *Recommendation 3: The AABB cannot recommend for or against a liberal or restrictive transfusion threshold for hospitalized, hemodynamically stable patients with the acute coronary syndrome.* (Grade: uncertain recommendation; very low-quality evidence.)
  2. Society of Thoracic Surgeons/ The Society of Cardiovascular Anesthesiologists: *With hemoglobin levels below*

6 g/dL, red blood cell transfusion is reasonable since this can be lifesaving. Transfusion is reasonable in most postoperative patients whose hemoglobin is less than 7 g/dL but no high level evidence supports this recommendation. (Level of evidence C. CLASS IIa - Additional studies with focused objectives needed. IT IS REASONABLE to perform procedure/administer treatment.)

3. Society of Critical Care Medicine: *"A "restrictive" strategy of RBC transfusion (transfusion when Hb <7 g/dL) is as effective as a "liberal" strategy (transfusion when Hb < 10 g/dL) in critically ill patients with hemodynamically stable anemia, except possibly in patients with acute myocardial ischemia". (*Grade-Level 1. The recommendation is convincingly justifiable based on the available scientific information alone.)

- Findings from 2 systematic reviews included a Cochrane Review that addressed evidence for the effect of transfusion thresholds on the use of allogeneic and/or autologous red cell transfusion, and the evidence for any effect on clinical outcomes. The developer also presented a Salpeter Meta-Analysis and Systematic Review that looked at randomized controlled trials evaluating a restrictive transfusion trigger of <7 g/dL.

Guidance from the Evidence Algorithm

Process measure based on SR/grading (Box 3) → QQC provided (Box 4) → Moderate quality evidence (Box 5b) → MODERATE

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured?*

Preliminary rating for evidence: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

Although there is no performance data on the measure as specified, the developer provided data on blood transfusion appropriateness and rate of hospitalization with blood transfusion that indicates opportunity for improvement.

Disparities

The developed indicated that no disparity data are available.

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are data available to show the percent of transfusions when hemoglobin levels are at the various lower levels specified?*
- *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus

- I think this is too hard a concept to measure. Has the strong unintended consequence of delaying transfusion in patients who need it. Also does not account for the vector (how fast is bleeding?) and the need to make decisions in the fog of war.
  Evidence is more controversial than stated.
  Experts moving away from absolute threshold and towards "is patient bleeding?" and "how are they doing?"

- This is a Trial Approval Program process and outcome eMeasure for restrictive transfusion strategy.

  There are 2 Cochrane systematic reviews specific to this measure; quality, quantity and consistency of the evidence has been provided; the evidence is graded; and is supported by clinical guideline recommendations from the AABB, STS/Society of CV Anesthesiologists and Society of Critical Care Medicine.

1b. Performance Gap

- No performance or disparity data was provided.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): EHR

Specifications: HQMF specifications are provided – see technical review

- Numerator Statement: Patients whose hemoglobin level measured prior to the transfusion and closest to the transfusion was:
  - less than 7.0 grams
  - >=7.0 and <8.0 grams
  - >=8.0 and <9.0 grams
  - >=9.0 and <10.0 grams
  - 10.0 grams or greater
- Denominator Statement: Patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion
- Denominator Exclusions:
  - Patients who have a surgical procedure performed to address a traumatic injury
  - Patients who have a solid organ transplant
  - Patients undergoing extracorporeal membrane oxygenation (ECMO) treatment at the time of initial transfusion.
  - Patients whose first unit of whole blood or packed red blood cells was given while an Emergency Department patient.
  - Patients with sickle cell disease or hereditary hemoglobinopathy

eMeasure Technical Advisor(s) review:

Submitted measure is an HQMF compliant eMeasure	The submitted eMeasure specifications follow the industry accepted format for eMeasure (HL7 Health Quality Measures Format (HQMF)). **HQMF specifications** ☒ **Yes** ☐ **No**
Documentation of HQMF or QDM limitations	N/A – All components in the measure logic of the submitted eMeasure are represented using the HQMF and QDM
Value Sets	The submitted eMeasure specifications uses existing value sets when possible and uses new value sets that have been vetted through the VSAC
Measure logic is unambiguous	Submission includes test results from a simulated data set demonstrating the measure logic can be interpreted precisely and unambiguously;

	Submitted with Bonnie results
Feasibility Testing	The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval.

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

Initial reliability testing was conducted in the Bonnie test deck; the overall patient simulation included 48 patients. The developer stated that Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. As a measure under consideration for the Trial Approval program, the developers must indicate if they have a plan in place for full testing (reliability and validity) and this information will be submitted and evaluated by NQF prior to any consideration of full measure endorsement. The Testing attachment indicates a plan for reliability and validity testing.

Questions for the Committee:

- *The Committee will not be asked to vote on Reliability for this eMeasure since it is being considered for Trial Use; however, questions regarding the testing plan and other concerns about reliability are welcome for discussion.*

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Based on the information provided, and intent of the measure, do you feel the specifications are consistent with evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

The only testing completed to date includes Bonnie testing and some review for feasibility. Additionally, the developer stated that findings from public comment support the face validity of this measure. The public comment was open for 30 days and the Joint Commission received 150 responses. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5=Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.41
Denominator clearly describes the activity being measured	4.40
Numerator inclusions clear and appropriate	4.46
Denominator inclusions clear and appropriate	4.42
Numerator exclusions clear and appropriate	4.44

Denominator exclusions clear and appropriate	4.36
Accurately assesses the process of care to which it is addressed	4.31

This measure is being considered for trial use, thus full validity testing results are not expected and the Committee will not vote on this criterion.

2b3-2b7. Threats to Validity

2b3. Exclusions:
When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Patients with a traumatic injury <=48 hours prior to or during the encounter.
- Patients with a solid organ transplant <=48 hours prior to or during the encounter.
- Patients who have an ECMO procedure during the inpatient encounter.
- Patients with sickle cell disease and related blood disorders

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

Unknown at this time.

2b6. Comparability of data sources/methods:
N/A

2b7. Missing Data

The developer stated that eMeasures are calculated using only the structured data collected in certified EHR technology. Data not present in the structured field from which the measure draws will not be included in the measure calculation.

The Committee will only vote on one portion of Scientific Acceptability: 2b1 – to determine if the measure specifications are consistent with evidence. This is a must pass criteria.

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1. & 2b1. Specifications

- The specifications are consistent with the evidence, though full testing of the eMeasure, as specified has not been conducted to assess reliability or validity.

2b2. Validity – Testing

- In addition to Bonnie testing, TJC held a 30 day public comment period with 150 responses that support the face validity of the measure, per the developer.

  Full reliability and validity testing must be completed prior to NQF consideration of full measure endorsement.

2b3-7. Threats to Validity

- Data not present in an EHR structured field for the measure will not be included in the measure calculation.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The feasibility analysis submitted by the measure developer meets the requirements to be considered for eMeasure Trial Approval. Based on the findings of the eMeasure Technical Review, the submitted eMeasure specification is capable of being processed and interpreted by clinical information systems and is ready for implementation in real world settings.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*
- *Is the data collection strategy ready to be put into operational use?*
- *Does the eMeasure Feasibility Score Card demonstrate acceptable feasibility in multiple EHR systems and sites?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

3. Feasibility

- Following feasibility analysis the measure meets the requirements to be considered for eMeasure Trial Approval.

  Data source is the EHR. Not all facilities have an EHR. The denominator exclusions may not be complete. Abstraction of the chart may require significant modification of the data fields in the EHR or addition of staff to abstract the paper record.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification.

Improvement results N/A

Unexpected findings (positive or negative) during implementation N/A

Potential harms None identified

Feedback :

None identified

***Questions for the Committee**:*

- *Does the Committee consider the certification program in Blood Management to be an accountability program?*
- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

4. Usability and Use

- This measure is not publically reported. Planned to be used in an accountability program. TJC maintains a voluntary Blood Management certification program.

Criterion 5: Related and Competing Measures

Related or competing measures

N/A

Harmonization

N/A

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: PBM-04: Initial Transfusion Threshold

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 5/20/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to

demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Assessment of the proportion of various pre-transfusion hemoglobin levels in patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion.☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

3. Process: Assessment of pre-transfusion hemoglobin level, and administration of transfusion if hemoglobin below 7 or 8, or if clinical presentation requires transfusion at higher hemoglobin levels (restrictive strategy).
4. Outcomes: A. Avoidance of transfusion and attendant complications. B. Reduced length of stay, reduced morbidity and mortality. C. Blood resource conservation.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☒ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

A. AABB

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Carson JL, Grossman BJ, Kleinman S, Tinmouth at, et al. Red Blood Cell Transfusion: A Clinical Practice Guideline From the AABB. *Ann Intern Med.* 2012;157(1):49-58.

http://annals.org/article.aspx?articleid=1206681

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Recommendation 1: The AABB recommends adhering to a restrictive transfusion strategy (7 to 8 g/dL) in hospitalized, stable patients.

Recommendation 2: The AABB suggests adhering to a restrictive strategy in hospitalized patients with preexisting cardiovascular disease and considering transfusion for patients with symptoms or a hemoglobin level of 8 g/dL or less.

Recommendation 3: The AABB cannot recommend for or against a liberal or restrictive transfusion threshold for hospitalized, hemodynamically stable patients with the acute coronary syndrome.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Recommendation 1: Grade: strong recommendation; high-quality evidence.

Recommendation 2: Grade: weak recommendation; moderate-quality evidence.

Recommendation 3: Grade: uncertain recommendation; very low-quality evidence.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

The strength of recommendations (for or against intervention) is graded as "strong" (indicating judgment that most well-informed people will make the same choice; "We recommend . . . "), "weak" (indicating judgment that a majority of well-informed people will make the same choice, but a substantial minority will not; "We suggest . . . "), or "uncertain" (indicating that the panel made no specific recommendation for or against interventions; "We cannot recommend . . . ").

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☒ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

B. Society of Thoracic Surgeons/ The Society of Cardiovascular Anesthesiologists

1a.4.1. Guideline citation (*including date*) and URL for guideline (*if available online*):

Ferraris V, Brown JR, Despotis GJ, Hammon JW, et al. 2011 Update to the Society of Thoracic Surgeons and the Society of Cardiovascular Anesthesiologists Blood Conservation Clinical Practice Guidelines.

Ann Thorac Surg 2011;91:944-82.

http://www.annalsthoracicsurgery.org/article/S0003-4975(10)02888-2/pdf

1a.4.2. Identify guideline recommendation number and/or page number and quote verbatim, the specific guideline recommendation.

(No Number Table 2): "With hemoglobin levels below 6 g/dL, red blood cell transfusion is reasonable since this can be lifesaving. Transfusion is reasonable in most postoperative patients whose hemoglobin is less than 7 g/dL but no high level evidence supports this recommendation. (Level of evidence C)."

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Class IIA **CLASS IIa,** *Benefit >> Risk*
Additional studies with focused objectives needed. **IT IS REASONABLE** to perform procedure/administer treatment.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system.
(*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

CLASS I, *Benefit >>> Risk*
Procedure/Treatment **SHOULD** be performed/administered

CLASS IIb, *Benefit ≥ Risk*
Additional studies with broad objectives needed; additional registry data would be helpful.
Procedure/Treatment **MAY BE CONSIDERED**

CLASS III, *Risk ≥ Benefit*
Procedure/Treatment should **NOT** be performed/administered **SINCE IT IS NOT HELPFUL AND MAY BE HARMFUL**

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*):

Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines

http://professional.heart.org/idc/groups/ahamah-public/@wcm/@sop/documents/downloadable/ucm_319826.pdf

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → *complete section 1a.7*

☒ No → *report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7*

C. American Red Cross Guideline

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Vassallo R, et al. A Compendium of Transfusion Practice Guidelines, Second Edition, 2013. American Red Cross, page 8.
http://www.redcrossblood.org/sites/arc/files/59802_compendium_brochure_v_6_10_9_13.pdf

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page 15: A restrictive RBC transfusion strategy (Hgb 7–8 g/dL trigger) is recommended in stable hospitalized patients.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

No grade assignment

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

n/a

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

n/a

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

D. Society of Critical Care Medicine:

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

Napolitano L, et al. Clinical Practice Guideline: Red blood cell transfusion in adult trauma and critical care. Crit Care Med 2009 Vol 37, No 12.

http://journals.lww.com/ccmjournal/Abstract/2009/12000/Clinical_practice_guideline__Red_blood_cell.19.aspx

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

Page 3127, recommendation 3: "A "restrictive" strategy of RBC transfusion (transfusion when Hb <7 g/dL) is as effective as a "liberal" strategy (transfusion when Hb < 10 g/dL) in critically ill patients with hemodynamically stable anemia, except possibly in patients with acute myocardial ischemia".

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

Level 1. The recommendation is convincingly justifiable based on the available scientific information alone. This recommendation is usually based on Class I data, however strong Class II evidence may form the basis for a Class 1 recommendation.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

Level 2. The recommendation is reasonably justifiable by available scientific evidence and strongly supported by expert opinion. This recommendation is usually supported by Class II data or a preponderance of Class III evidence.

Level 3. The recommendation is supported by available data but adequate scientific evidence is lacking.

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

Same, p. 3126.

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

A. Cochrane Review

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Carson JL, Carless PA, Hebert PC. Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion (review). *The Cochrane Collaboration*, April 2012.

http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD002042.pub3/full

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Same as for evidence review (see full text article); no grading applied.

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

C. Salpeter Meta-Analysis and Systematic Review:

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

Salpeter SR, Buckley JS, Chatterjee s> Impact of More Restrictive Blood Transfusion Strategies on Clinical Outcomes: A Meta-analysis and Systematic Review. *The American Journal of Medicine*, Vol 127, No 2, February 2014.

http://www.ncbi.nlm.nih.gov/pubmed/24331453

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Same as for evidence review (see full text article); no grading applied.

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

A. **Cochrane Review:**

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Evidence for the effect of transfusion thresholds on the use of allogeneic and/or autologous red cell transfusion, and the evidence for any effect on clinical outcomes.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

No grade assigned.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

n/a

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).
Date range: "Unrestricted"

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

19 controlled trials were ultimately included, covering a span of 55 years and 6,264 patients.

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Selection bias – Low for 9 trials, insufficient information for 9 trials, high risk for 1 trial

Allocation bias – low for 4 trials, unclear for 13 trials, high risk for 2 trials

Blinding of physicians – low risk for all trials

Incomplete outcome data – low risk for 14 trials, unclear for 5 trials

Selective reporting – none

Eighteen of nineteen trials presented data suitable for inclusion in the meta-analysis.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Results were that "In patients who do not have acute coronary artery disease, blood transfusion can probably be withheld in the presence of haemoglobin levels as low as 7.0 g/dL to 8.0 g/dL as long as there is no notable bleeding." In addition, "…restrictive transfusion strategies were associated with a reduction of more than one-third in the number of patients receiving blood, a red cell transfusion requirement that was approximately one unit lower, and a haemoglobin concentration (average postoperative) that was around 1.5 g/dL lower than in the blood transfusion group."

Ratings of statistical significance and confidence intervals varied by studied outcome; heterogeneity between studies was statistically significant (P<0.00001, Chi 96.82).

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

"None of the outcomes evaluated, including mortality, cardiac morbidity, infections, and length of hospital stay, appear to be adversely affected by the lower use of red cell transfusions."

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

n/a

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

B. **Clinical Practice Guideline, AABB:**

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

The GRADE system (39) uses the following 4 ratings for quality of evidence:

"High" indicates considerable confidence in the estimate of effect. The true effect probably lies close to the estimated effect, and future research is unlikely to change the estimate of the health intervention's effect.

"Moderate" indicates confidence that the estimate is close to the truth. Further research is likely to have an important effect on confidence in the estimate and may change the estimate of the

health intervention's effect.

"Low" indicates that confidence in the effect is limited. The true effect may differ substantially from the estimate, and further research is likely to have an important effect on confidence in the estimate of the effect and is likely to change the estimate.

"Very low" indicates little confidence in the effect estimate. Any estimate of effect is very uncertain.

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

C. **Salpeter Meta-Analysis and Systematic Review:**

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

Randomized controlled trials evaluating a restrictive transfusion trigger of <7 g/dL.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

No grade assigned.

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

n/a

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*).

Date range: 1966 to April 2013.

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

Primary analysis – 3 RCTs (trigger <7 g/dL) Secondary analysis – 19 RCTS (trigger 7.5 – 10 g.dL).

1a.7.6. What is the overall quality of evidence <u>across studies</u> in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Unstated

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) <u>across studies</u> in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

In the primary analysis, pooled results from 3 trials with 2364 participants showed that a restrictive hemoglobin transfusion trigger of <7 g/dL resulted in reduced in-hospital mortality (risk ratio [RR], 0.74; confidence interval [CI], 0.60-0.92), total mortality (RR, 0.80; CI, 0.65-0.98), rebleeding (RR, 0.64; CI, 0.45-0.90), acute coronary syndrome (RR, 0.44; CI, 0.22-0.89), pulmonary edema (RR, 0.48; CI, 0.33-0.72), and bacterial infections (RR, 0.86; CI, 0.73-1.00), compared with a more liberal strategy. Pooled data from randomized trials with less restrictive transfusion strategies showed no significant effect on outcomes.

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

In-hospital mortality, total mortality, ACS, pulmonary edema, and infection occurrence were all assessed, with reduction in occurrence of all noted in a more restrictive strategy.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for <u>each</u> new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

n/a

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

In January 2015 a literature search of EMBASE, Pub Med, MEDLINE and other relevant sources including professional association websites, The Cochrane Library, the National Guideline Clearinghouse, and other sources was conducted, using search terms such as anemia, preoperative testing, and other relevant search terms, requesting English language publications from 2009 – 2014. Identified publications were searched for additional relevant reference documents.

1a.8.2. Provide the citation and summary for each piece of evidence.

New York State Department of Health:

"Hemoglobin concentration:

Hb >10 g/dL– Transfusion is rarely indicated

Hb 6-10 g/dL – indications for transfusion should be based on the patient's risk of inadequate oxygenation from ongoing bleeding and/or high-risk factors, such as age, cardiovascular compromise, or respiratory disease.

Hb <6 g/dL – transfusion is almost always indicated"

New York State Council on Human Blood and Transfusion Services. Guidelines for Transfusion Options and Alternatives, 2010. Downloaded from www.wadswoth.org/labcert/blood_tissue July 2015.

American Academy of Family Physicians: "The threshold for transfusion of red blood cells should be a hemoglobin level of 7g/dL (70 g/L) in most adults and children." Evidence Rating A, RCTs in adults and children with critical illness. Transfusion of Blood and Blood Products: Indications and Complications. Am Fam Physician 2011;83(6): 719-24.

Sharma S, Sharma P, Tyler L. Transfusion of Blood and Blood Products: Indications and Complications. *American Family Physician*. March 15, 2011: volume 83 No. 6, p.720.
"The threshold for transfusion of red blood cells should be a hemoglobin level of 7 g per dL (70 g per L) in adults and most children." (Evidence rating A).

Shander et al. Appropriateness of Allogeneic Red Blood Cell Transfusion: The International Consensus Conference on Transfusion Outcomes. *Transfusion Medicine Reviews*, Vol 25, No 3 (July), 2011: pp 232-246.e53.
An international multidisciplinary panel of 15 experts reviewed 494 published articles and used the RAND/UCLA Appropriateness Method to determine the appropriateness of allogeneic red blood cell (RBC) transfusion based on its expected impact on outcomes of stable nonbleeding patients in 450 typical inpatient medical, surgical, or trauma scenarios. Panelists rated allogeneic RBC transfusion as appropriate in 53 of the scenarios (11.8%), inappropriate in 267 (59.3%), and uncertain in 130 (28.9%). Red blood cell transfusion was most often rated appropriate (81%) in scenarios featuring patients with hemoglobin (Hb) level 7.9 g/dL or less, associated comorbidities, and age older than 65 years. Red blood cell transfusion was rated inappropriate in all scenarios featuring patients with Hb level 10 g/dL or more and in 71.3% of scenarios featuring patients with Hb level 8 to 9.9 g/dL. Conversely, no scenario with patient's Hb level of 8 g/dL or more was rated as appropriate. Nearly one third of all scenarios were rated uncertain, indicating the need for more research. The observation that allogeneic RBC transfusions were rated as either inappropriate or uncertain in most scenarios in this study supports a more judicious transfusion strategy.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
PBM_04_evidence_attachment.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
All published sources indicate that a strict transfusion strategy is preferable to a liberal strategy, since transfusion can be harmful and contributes to higher mortality, infection, and other complications.1,2,3,4 Most guidelines recommend a threshold of 7.0 or 8.0 grams of hemoglobin or less as an indication for transfusion, and if the hemoglobin level is 10.0 or greater there is agreement that the transfusion is rarely indicated. There is disagreement among guidelines, however, when patients have underlying cardiac disease, postoperative status, or other clinical conditions5, others, and there is some concern about a lack of robust evidence to support some guidelines.6,7 Caution is always advised that when determining the appropriateness of transfusion, underlying clinical conditions and symptoms should be taken into consideration.
The purpose of this measure is to allow facilities to profile blood usage according to initial transfusion hemoglobin thresholds. Strata are defined to direct facility review to the appropriateness of selected transfusions, taking into account clinical symptoms combined with hemoglobin measurements. By this review, facilities will be able to determine the best approaches to enhance blood conservation and management and over time, there should be a gradual decline in the proportion of initial units given from the higher hemoglobin values to those lower values supported in a restrictive transfusion strategy in the literature and guidelines as part of a blood management program.

1. Carson JL, Grossman BJ, Kleinman S, Tinmouth at, et al. Red Blood Cell Transfusion: A Clinical Practice Guideline From the AABB. Ann Intern Med. 2012;157(1):49-58.
2. Goodnough LT, Shander A. Patient Blood Management. Anesthesiology v116; No 6, June 2012
3. Paone G, Likosky DS, Brewer R, Theurer PF, et al. Transfusion of 1 and 2 Units of Red Blood Cells is Associated With Increased Morbidity and Mortality. Ann Thorac Surg 2014; 97:87-94.
4. Shander A, Goodnough LT. Can Blood Transfusion Be Not Only Ineffective, But Also Injurious? Ann Thorac Surg 2014; 97: 11-4.
5. Shander A, Gross I, Hill S, Javidroozi M, Sledge S. A new perspective on best transfusion practices. Blood Transfus 2013; 11: 193-202.
6. Carson JL, Carless PA, Hebert PC. Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion (review). The Cochrane Collaboration, April 2012.
7. Wilkinson KL, Brunskill SJ, Doree C, Hopewell S, et al.The Clinical Effects of Red Blood Cell Transfusions: An Overview of the Randomized Controlled Trials evidence Base. Transfusion Medicine Reviews, Vol 25, No.2 (April), 2011, pp 145-155 e2.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
This is a new measure for which approval for trial use is requested.

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
A. Agency for Healthcare Research and Quality (AHRQ): Blood transfusion was the most common listed procedure performed during hospitalizations in 2010 (11 percent of stays with a procedure); the rate of hospitalization with blood transfusion has more than doubled since 1997. The percentage change in rate of all stays with a blood transfusion from 1997 – 2010 is 126%. Most Frequent Procedures Performed in U.S. Hospitals, 2010. Healthcare Cost and Utilization project, statistical brief, February 2013, AHRQ.

B. An international multidisciplinary panel of 15 experts reviewed 494 published articles and used the RAND/UCLA Appropriateness Method to determine the appropriateness of allogeneic red blood cell (RBC) transfusion based on its expected impact on outcomes of stable nonbleeding patients in 450 typical inpatient medical, surgical, or trauma scenarios. Panelists rated allogeneic RBC transfusion as appropriate in 53 of the scenarios (11.8%), inappropriate in 267 (59.3%), and uncertain in 130 (28.9%). Red blood cell transfusion was most often rated appropriate (81%) in scenarios featuring patients with hemoglobin (Hb) level 7.9 g/dL or less, associated comorbidities, and age older than 65 years. Red blood cell transfusion was rated inappropriate in all scenarios featuring patients with Hb level 10 g/dL or more and in 71.3% of scenarios featuring patients with Hb level 8 to 9.9 g/dL. Conversely, no scenario with patient´s Hb level of 8 g/dL or more was rated as appropriate. Nearly one third of all scenarios were rated uncertain, indicating the need for more research. The observation that allogeneic RBC transfusions were rated as either inappropriate or uncertain in most scenarios in this study supports a more judicious transfusion strategy. Appropriateness of Allogeneic Red Blood Cell transfusion: The International Consensus Conference on Transfusion Outcomes. Transfusion Medicine Reviews, Vol 25, No. 3 (July)

2011.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
No disparities were identified.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
No disparity data was found in the literature.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, Frequently performed procedure, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
1. Agency for Healthcare Research and Quality (AHRQ): Blood transfusion was the most common of all listed procedures performed during hospitalizations in 2010 (11 percent of stays with a procedure); the rate of hospitalization with blood transfusion has more than doubled since 1997. The percentage change in rate of all stays with a blood transfusion from 1997 – 2010 is 126%.
2. AABB: If a restrictive transfusion strategy were widely implemented and replaced a liberal strategy, exposure of patients to red blood cell (RBC) transfusions would decrease by an average of approximately 40% (relative risk, 0.61 [confidence interval (CI), 0.52 to 0.72]). This would have a large effect on blood use and the risks for infectious and noninfectious complications of transfusion. Unnecessary transfusions increase costs and expose patients to potential infectious or noninfectious risks, such as hepatitis B and C virus, human immunodeficiency virus (HIV), transfusion-associated circulatory overload, transfusion-related acute lung injury, fatal hemolysis, life-threatening reaction, and fever.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**
1. Most Frequent Procedures Performed in U.S. Hospitals, 2010. Healthcare Cost and Utilization project, statistical brief, February 2013, AHRQ.
2. Red blood cell transfusion: a clinical practice guideline from the AABB. Ann Intern Med. 2012 Jul 3;157(1):49-58. [63 references]

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*
Not a PRO-PM.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across

organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. Subject/Topic Area *(check all the areas that apply):*
Surgery

De.6. Cross Cutting Areas *(check all the areas that apply):*
Overuse, Safety

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
https://www.jointcommission.org/measure_development_initiatives.aspx

S.2a. If this is an eMeasure, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is an eMeasure **Attachment:** PBM-04_InitialTransfusionThreshold.zip

S.2b. Data Dictionary, Code Table, or Value Sets *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
Attachment **Attachment:** InitialTransfusionThreshold_v4_3_Wed_Jun_08_10.20.18_CDT_2016.xls

S.3. For endorsement maintenance, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
n/a

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Patients whose hemoglobin level measured prior to the transfusion and closest to the transfusion was:
- less than 7.0 grams
- >=7.0 and <8.0 grams
- >=8.0 and <9.0 grams
- >=9.0 and <10.0 grams
- 10.0 grams or greater

S.5. Time Period for Data *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Episode of care

S.6. Numerator Details *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Hemoglobin level prior to and closest to the transfusion is represented by a code from the following Value Set and associated QDM datatype:
- "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma LOINC Value Set (2.16.840.1.113762.1.4.1104.4)

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
Patients age 18 and over receiving the first unit of a whole blood or packed cell transfusion

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

Inpatient encounters are represented by a code from the following value set and associated QDM datatype:
•"Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"

Patients who receive the first unit of a packed cell or whole blood transfusion are represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*

- Patients who have a surgical procedure performed to address a traumatic injury
- Patients who have a solid organ transplant
- Patients undergoing extracorporeal membrane oxygenation (ECMO) treatment at the time of initial transfusion.
- Patients whose first unit of whole blood or packed red blood cells was given while an Emergency Department patient.
- Patients with sickle cell disease or hereditary hemoglobinopathy

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*

Patients who have a surgical procedure performed to address a traumatic injury are represented by a code from the following Value Set and associated QDM datatype:
"Attribute: Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)
Patients who have a solid organ transplant are represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
Patients who undergo ECMO at the time of initial transfusion are represented by a code from the following Value Set and associated QDM datatype:
"Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)
Patients whose first unit is given while an Emergency Department patient are implicity excluded as blood administered in an ED location is not captured in this measure.

Patients with sickle cell disease or hereditary hemoglobinopathy are represented by a code from the following Value Set and associated QDM datatype:
Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*

Stratification 1 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Blood Transfusion Administration"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result < 7.0 g)"

Stratification 2 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Blood Transfusion Administration"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 7.0 g)
(result < 8.0 g)

Stratification 3 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Blood Transfusion Administration"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:

(result >= 8.0 g)
(result < 9.0 g)

Stratification 4 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Blood Transfusion Administration"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" satisfies all:
(result >= 9.0 g)
(result < 10.0 g)

Stratification 5 =
AND: Most Recent: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma" <= 45 day(s) starts before start of "Occurrence A of Procedure, Performed: Blood Transfusion Administration"
AND: "Occurrence A of Laboratory Test, Performed: Hemoglobin blood serum plasma (result >= 10.0 g)"

S.13. Risk Adjustment Type (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. Identify the statistical risk model method and variables *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
n/a

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
n/a

S.16. Type of score:
Count
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Score within a defined interval

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
See attached HQMF file.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Records are not sampled.

S.21. Survey/Patient-reported data *(If measure is based on a survey, provide instructions for conducting the survey and guidance on*

minimum response rate.)
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Measure is not based on a survey; not a PRO-PM.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
eMeasures are calculated using only the structured data collected in certified EHR technology (CEHRT). Data not present in the structured field from which the measure draws will not be included in the measure calculation.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Laboratory

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
Hospitals report EHR data using Certified Electronic Health Record Technology (CEHRT), and by submitting Quality Reporting Document Architecture Category 1 (QRDA-1).

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No data collection instrument provided

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite measure.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
PBM_04_testing_form_for_trial_use.docx,PBM04_CMS608v0_Bonnie_Export.xlsx

National Quality Forum
Measure Testing Form for Trial Approval Program

Measure Title: PBM-04: Initial Transfusion Threshold

Date of Submission: 5/31/2016

Type of Measure:

☐ Composite –	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

A measure submision that is to be considered for the Trial Approval Program must complete this form in its entirety. Either a test data set provided by the measure developer, or the use of the Bonnie tool is acceptable to provide prelminary testing results,

For all measures being submitted for potential acceptance into the Trial Approval Program, each section must be filled out as completely as possible.

Respond to all questions as instructed with answers immediately following the question. All information on testing of either a sample data set or results from Bonnie testing that can demonstrate, to the extent possible, the the measure meets the reliability and validity must be in this form..

If you are unable to check a box, please highlight or shade the box for your response.

Maximum of 10 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***

Contact NQF staff regarding questions at trialmeasures@qualityforum.org

DATA and SAMPLING INFORMATION

1. DATA/SAMPLE USED FOR PRELMINARY TESTING OF THIS MEASURE

It is important that the measure developer use a data set to conduct preliminary testing in order to evauate the measure logic and the inclusions/exclusions for the population used in the measure.

What type of data was used for testing? (*The measure developer must provide a test data set that will provide some initial information to be used for the evaluation, or the Bonnie testing tool can use can be used to create a sample data set using synthesized patients.)* Please indicate whether the test data set used was provided through the measure developer, or through the Bonnie tool.

> The Bonnie testing tool was used to simulate a testing environment where measure specifications and HQMF output are tested against synthetic test data. Measure developers rely on the results in Bonnie to confirm whether the measure logic is performing as expected.
>
> Reference the eCQI Resource Center website (https://ecqi.healthit.gov/ecqm-tools/tool-library/bonnie) or the Bonnie testing tool website (https://bonnie.healthit.gov/) for more information about Bonnie functionality and its role in measure development. Please also reference the Bonnie testing worksheet attachment for detailed Bonnie test cases and testing results for this measure.

If Bonnie was NOT used, please identify the specifications for the test dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured)*

Not Applicable

What levels of analysis were tested (either through the test data set or Bonnie)? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*) *in order to determine its suitability for inclusion into the Trial Approval Program.,*

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency

☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other:

1.4. How many and which <u>patients</u> were included in the testing and analysis (by level of analysis and data source)? (*Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis)*

48 unique synthetic patient records were created in the BONNIE testing system for this measure. Cases were used to test the validity of each data element and timing relationship in the measure. Bonnie testing was also performed for each stratum specified in the measure. Patient characteristics such as age, diagnosis, and length of stay were pre-determined to provide a variety of scenarios that adequately tested for patients passing each data element and failing each data element. Data included in cases and tested for this measure included all data elements required to calculate the measure and the measure denominator exclusions.

For further information on the characteristics of the patients included in the analysis, please refer to the attached BONNIE testing spreadsheet.

1.5. Please refer to the guidance for Bonnie testing found at this link. Bonnie testing results may be compiled into spreadsheet or table, which must be completed in its entirety, to the extent possible, in order to provide a basis for evaluation to determine the acceptability of the measure for inclusion in the Trial Approval program. *Any questions regarding the completion of this form can be directed to NQF Staff at trialmeasures@qualityforum.org.*

Please refer to the attached BONNIE testing spreadsheet.

<u>RELIABILITY AND VALIDITY ASSESSMENTS</u>

<u>Note</u>: The information provided in this next section is intended to aid the Standing Committee and other stakeholders in understanding to what degree the measure is both reliable and valid. While it is not possible to provide comprehensive results due to the lack of actual testing data, the developer needs to provide as much information as possible based on their interpretation of the results from the sample test data.

2.1 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 **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating reliability**? (i.*e., what do the sample results mean and what are the norms for the test conducted?*) Please summarize the plan for future testing of reliability if the measure is accepted into the Trial Approval Program. **Include descriptions of:**

Inter-abstractor reliability, and data element reliability of all critical data elements

Computation of the performance measure score (e.g., signal-to-noise analysis)?

All data elements within the measure are specified using nationally accepted standard terminologies, including LOINC, SNOMEDCT, ICD10CM, and ICD10PCS. Bonnie testing confirms that the measure logic performs as expected and that the terminologies used are applied consistently. This suggests that organizations using these terminologies within the EHR should be able to produce repeatable and reliable results. For further discussion of measure feasibility, please review the attached feasibility scorecard and feasibility report.

When data are available, The Joint Commission will perform extensive tests of measure reliability at the data element and measure level. Testing will include re-abstraction to the eCQM specification to evaluate missing data and assure inter-rater reliability, as well as analysis of agreement rates for data elements used to compute measure rates for PBM-04.

2.2 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 **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score. **What is your interpretation of the results in terms of demonstrating validity**? (i.*e., what do the results mean and what are the norms for the test conducted?*). Please summarize the plan for future testing of validity if the measure is accepted into the Trial Approval Program. Include the method(s) of validity testing and what it will test (describe the steps—do not just name a method; what will be tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis will be used used)

Face validity was established through public comment.

Public comment was open for 30 days from March 20 to April 19, 2015. The Joint Commission received 150 responses to the call for comment. Respondents were asked to rate the measure on a number of parameters, using a Likert scale ranging from 1 to 5, where 1=Disagree and 5-Agree. The table below presents the average rating for these parameters.

PARAMETER	RATING
Numerator clearly describes the activity being measured	4.41
Denominator clearly describes the activity being measured	4.40
Numerator inclusions clear and appropriate	4.46
Denominator inclusions clear and appropriate	4.42
Numerator exclusions clear and appropriate	4.44
Denominator exclusions clear and appropriate	4.36
Accurately assesses the process of care to which it is addressed	4.31

Findings from public comment support the face validity of this measure.

The Bonnie testing tool and environment were used to establish content and construct validity through testing of the measure logic and value sets. Each data element and logic statement was tested to confirm actual results met expectations. Bonnie testing includes negative and positive testing of each data element in the measure. Positive testing ensures patients expected to be included in the measure are included. Negative testing ensures that patients who do not meet the data criteria are not included in the measure. An example of negative testing would be to include test cases with pediatric ages to ensure that pediatric patients are not included in the measure.

Initial Population and Denominator test cases positively test to ensure that only patients >= 18 years of age who have a surgical procedure performed <=48 hours prior to the inpatient encounter or during the inpatient encounter are included. Negative test cases ensure that patients who do not meet these criteria to do not pass into the denominator. For example, cases test patients who have a surgical procedure at 49 hours and 48 hours prior to the start of the encounter. Patients who have a surgical procedure 48 hours

prior to the start of the encounter were included in the denominator, while patients with a surgical procedure at 49 hours prior to the encounter were not.

Numerator test cases positively test to ensure patients who have a hemoglobin result recorded <= 45 days and >=14 days prior to the start of surgery are included in the numerator. Negative test cases ensure that a patient who did not meet these criteria are not included. For example, test cases in which hemoglobin results were recorded >45 days prior to surgery or after surgery confirmed that such patients would not be included in the numerator.

Denominator exclusion test cases for this measure ensure that patients are properly removed from the denominator if they have specific documented procedures or encounter diagnoses. Negative test cases for the denominator exclusion ensure that patients without these diagnoses or procedures fall in to the denominator population. Testing confirmed patients meeting the exclusion criteria are removed from the measure appropriately, while those that do not meet the criteria are retained in the denominator population.

Once pilot data are available, The Joint Commission will evaluate construct validity though an examination of the degree of association between measure results for PBM-04 and other measures in this set, using the Pearson Correlation Coefficient. The Joint Commission would hypothesize that a relationship exists between this measure and other measures in the Patient Blood Management set.

In addition, data element validity would be assessed for accuracy and clarity in reliability testing, using the data element values obtained in the reliability study as the gold standard.

2.3 Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion. **What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis).* Please summarize the plan for future testing of exclusions if the measure is accepted into the Trial Approval Program. Describe the method of testing exclusions and what it will test (describe the steps—do not just name a method; what will be tested, e.g., whether exclusions affect overall performance scores; what statistical analysis will be used)

When data are available, The Joint Commission will analyze exclusion frequency and variability across providers. These data elements to be analyzed include:

- Patients with a traumatic injury <=48 hours prior to or during the encounter.
- Patients with a solid organ transplant <=48 hours prior to or during the encounter.
- Patients who have an ECMO procedure during the inpatient encounter.
- Patients with sickle cell disease and related blood disorders

2.4 Risk Stratification (applicable ONLY to outcome or resource use measures). If an outcome or resource use measure will not be risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities. If risk adjustment/stratification is needed then please describe the conceptual/clinical and statistical methods and criteria that will be used to select patient factors (clinical factors or sociodemographic factors) that will be used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not Applicable, not an outcome measure.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic health records (EHRs)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

Attachment **Attachment:** PBM04_NQF_Measure_Feasibility_Assessment_Report.docx

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

No modifications have been made

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Value sets are housed in the Value Set Authority Center (VSAC), which is provided by the National Library of Medicine (NLM), in coordination with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services.

Viewing or downloading value sets requires a free Unified Medical Language System® (UMLS) Metathesaurus License, due to usage restrictions on some of the codes included in the value sets. Individuals interested in accessing value set content can request a UMLS license at (https://uts.nlm.nih.gov/license.html)

There are no other fees or licensing requirements to use the Joint Commission performance measures, all of which are in the public

domain.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Regulatory and Accreditation Programs Professional Certification or Recognition Program Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

n/a

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new measure for which approval for trial use is requested.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

The Joint Commission maintains a certification program in Blood Management, which is a voluntary program for hospitals to achieve excellence in patient blood management. The measures in this set can be made available within a year for hospitals to use in fulfilling the requirements for certification. Hospitals using these measures evaluate care by these measures and submit data quarterly, either directly to The Joint Commission or through a vendor. The Joint Commission then generates reports and feeds the reports back to the certified organizations.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance

results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

n/a

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
n/a

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
There were no unintended negative consequences identified during testing.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
No

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
n/a

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
n/a

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.

No appendix **Attachment:**

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Joint Commission
Co.2 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-
Co.3 **Measure Developer if different from Measure Steward:** The Joint Commission
Co.4 **Point of Contact:** Tricia, Elliott, telliott@jointcommission.org, 630-792-5643-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
The role of the Technical Advisory Panel was to provide advisory oversight in literature review, measure construct and content, review of testing results, and endorsement of draft and finalized measures, as well as to continue to provide measure content oversight and update in the future.

eCQM Blood Management Technical Advisory Panel Member List
Richard J. Benjamin, MD, PhD, FRCPath, MS
Chief Medical Officer, Biomedical Services
American Red Cross, National Headquarters
7/15/15:
Chief Medical Officer
Cerus Corporation Laurence Bilfield, MD
Orthopaedic Surgeon
Cleveland Clinic HS - Lutheran
Lawrence Tim Goodnough, MD
Director, Transfusion Service Stanford Medical Center
Associate Director, Stanford Blood Center
Stanford University Medical Center Joseph E. Kiss, MD
Associate Professor of Medicine; Dept. of Medicine; Div. of Hem/Onc
Medical Director, Hemapheresis and Blood Services, CBB/ITxM
The Institute for Transfusion Medicine
University of Pittsburgh
Harvey G. Klein, MD
Senior Investigator
Transfusion Medicine Department
National Institutes of Health

Vijay K. Maker, MD, FACCS
Chairman, Department of Surgery
Executive Director, MGH Residency in General Surgery
Advocate Illinois Masonic Hospital
John (Jeffrey) McCullough, MD
Professor, Clinical Pathology, Blood Banking
University of Minnesota
Steven Frank, MD
Medical Director, The Johns Hopkins Health System
Blood Management Program
Associate Professor, Johns Hopkins Hospital,
Department of Anesthesiology and Critical Care
Medicine, Division of Vascular, Thoracic, Transplant Anesthesia
Neil K. Shah, M.D.
Medical Director of Informatics for Transfusion Services
Medical Director of Referral (Send Out) Testing
Stanford University Medical Center
Aryeh Shander, MD, FCCM, FCCP
Executive Medical Director of The Institute for Patient Blood Management and Bloodless Medicine and Surgery
Englewood Hospital and Medical Center
Jonathan H. Waters, MD, Chair
Medical Director in the Blood Management Division of Procirca, Inc.
Chief and Professor
Magee Women's Hospital
University of Pittsburgh

The purpose of the eCQM Task Force is to engage eCQM implementers in the electronic specification process, in order to produce clear, implementable eCQM specifications. Task force membership includes both hospital and vendor representatives with expertise in clinical informatics, electronic health record (EHR) implementation, and standard terminologies, as well as content experts with experience leveraging the EHR for blood management.

ePBM Task Force Roster

Irwin Gross, MD
Medical Director of Transfusion Services
Eastern Maine Medical Center
Hugh H. Ryan, MD
Senior Director & Chief Medical Officer
Population Health Programs
Cerner Corporation

Kimberly Bodine, DNP, RN
EHR Manager, Clinical Quality Measures and Clinical Analytics
Health Corporation of America
Douglas Van Deale, MD, FACS
Chief Medical Information Officer
University of Iowa

Jason Kratz, PhD
Inpatient eCQM Development Lead
Business Intelligence Developer
Epic
Cathy Bickerstaff, RN-BC
Informatics Specialist
St. Jude's Children's Research Hospital

Andrew Higgins, RN
Patient Blood Management Coordinator
Mayo Clinic

Catherine A Shipp, RN
Transfusion Safety Officer
Loyola University Medical Center
David Krusch, MD
Chief Medical Information Officer
Professor of Surgery
University of Rochester Medical Center
Lisa Gulker, DNP, ACNP-BC
Senior Director, Applied Clinical Informatics
Tenet Healthcare

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2016

Ad.3 **Month and Year of most recent revision:** 05, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annual

Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.7 **Disclaimers:** These performance measures are not clinical guidelines and do not establish a standard of medical care, and have not been tested for all potential applications. The measures and specifications are provided without warranty.

Ad.8 **Additional Information/Comments:**

NQF Measure Feasibility Assessment Report

Measure Title: PBM-04: Initial Transfusion Threshold

Measure Background and Overall Assessment of Measure Logic and Feasibility

The following assessment is conducted solely by the measure developer, The Joint Commission, based on our experience working with clinical experts, EHR and technical experts, and hospitals to assess feasibility throughout the measure development process. The measure was evaluated by five volunteer hospitals throughout the country during the fourth quarter of 2015.

This detailed report will provide a narrative summary of data elements found to be highly feasible, and will include verbatim scorecard responses for those data elements that were deemed to be more difficult to capture or for which there was great variability in feasibility. For complete scorecard results, please refer to the scorecard excel files attached to this submission.

Data Elements used in this Measure (in QDM format):

1. "Encounter, Performed: Encounter Inpatient" using "Encounter Inpatient SNOMEDCT Value Set (2.16.840.1.113883.3.666.5.307)"
2. "Laboratory Test, Performed: Hemoglobin blood serum plasma" using "Hemoglobin blood serum plasma Grouping Value Set (2.16.840.1.113762.1.4.1104.4)"
3. "Procedure, Performed: Blood Transfusion Administration" using "Blood Transfusion Administration SNOMEDCT Value Set (2.16.840.1.113762.1.4.1029.24)"
4. "Procedure, Performed: ECMO" using "ECMO Grouping Value Set (2.16.840.1.113762.1.4.1029.22)"
5. "Procedure, Performed: Solid Organ Transplant" using "Solid Organ Transplant Grouping Value Set (2.16.840.1.113762.1.4.1029.11)"
6. Attribute: "Diagnosis: Traumatic Injury" using "Traumatic Injury Grouping Value Set (2.16.840.1.113762.1.4.1029.10)"
7. Attribute: "Diagnosis: Sickle Cell Disease and Related Blood Disorders" using "Sickle Cell Disease and Related Blood Disorders Grouping Value Set (2.16.840.1.113762.1.4.1029.35)"

Initial Population and Denominator Data Elements

Data elements 1- "Encounter, Performed: Encounter Inpatient" and 3- "Procedure, Performed: Blood Transfusion Administration" are used to define the initial population and denominator of this measure.

On the feasibility scorecard, hospitals rated these data elements 1 and 3 as highly feasible when considering workflow, data availability, accuracy, definition, and use of standards.

Four out of five hospitals rated capture of data element 1 as highly feasible, represented as a score of 3 out of 3, for all domains of feasibility in both the current state and in the future. One site was not certain whether the data source for this data element was currently interfaced with the certified electronic health record. This site scored feasibility as a 1 for all domains in the current state, but as a 3 for future state, acknowledging that future state would be achieved much more quickly than the 3-5 year timeframe outlined in the scorecard, as the site would be interfacing this data in 2016 in order to report eCQMs.

Three out of five sites found data element 3 to be highly feasible in all domains except Data Standards- these hospitals have structured data fields for capture of transfusion data, but do not have those fields encoded in the terminology standard used in this measure, SNOMEDCT. These 3 sites reported that data capture would be highly feasible in the near term, stating that mapping this field to SNOMED would not be difficult and could be accomplished rather quickly. These three sites rated future state feasibility 3 out of 3, stating that capturing blood products in SNOMEDCT could occur in a much shorter timeframe than 3-5 years.

Two sites placed orders for blood in the EHR, but recorded blood product administration on paper. Both sites had plans to move to EHR-based barcode blood product administration, and found data element 3 to be highly feasible in the future state.

Numerator Data Element

Data element 2- "Laboratory Test, Performed: Hemoglobin blood serum plasma" is used to define the numerator for this measure. While some measures in this set require hemoglobin results recorded prior to the start of the encounter, PBM-04 evaluates hemoglobin results recorded within 45 days of the first blood transfusion. All sites have policies and practices in place that require a hemoglobin result prior to transfusion, and thus found this data element to be highly feasible, represented as a score of 3 out of 3 for all domains of feasibility.

Denominator Exclusions Data Elements

Data elements 4, 5, 6, and 7 are used to represent denominator exclusions.

Data element 4- "Procedure, Performed: ECMO," was found to be highly feasible by sites that perform ECMO. One site, a regional hospital, reported frequently using ECMO as a bridge for transport for patients requiring a higher level of care.

Data elements 6- "Diagnosis: Traumatic Injury," and 7- "Diagnosis: Sickle Cell Disease and Related Blood Disorders" both represent encounter diagnoses. All hospitals rated these data elements as highly feasible. Discussion around these data elements suggested that while missing data may occur due to clinician practice related to updating the patient problem list, the functionality to support collection of this data element is well established.

Feasibility for data element 5- "Procedure, Performed: Solid Organ Transplant" was found to be comparable to data element 6- "Procedure, Performed: Selected Elective Surgical Procedures." These data elements are found in the surgical schedule or operative record, and thus findings were similar, with the exception of sites that do not perform organ transplant, which would not use this data element.

Conclusion

Hospitals completing the feasibility scorecard reported the data elements required to calculate this measure to be highly feasibility in the current state. Of the measures in the PBM set, this measure received the highest ratings for feasibility. Approval for Trial Use status will support The Joint Commissions' efforts to further test this measure.

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3024
Measure Title: Carotid Endarterectomy: Evaluation of Vital Status and NIH Stroke Scale at Follow Up
Measure Steward: American College of Cardiology
Brief Description of Measure: Proportion of patients with carotid endarterectomy procedures who had follow up performed for evaluation of vital status and neurological assessment with an NIH Stroke Scale (by an examiner who is certified by the American Stroke Association)
Developer Rationale: There is a sound clinical rationale for systematically measuring the outcomes of carotid revascularization. First, without knowing the outcomes, a hospital cannot know if it is applying its treatment in a safe and effective manner. Given how infrequently current providers assess the 30-day survival and stroke outcomes, it is obvious that more than half of these hospitals have no foundation with which to assess the quality of their care.

Numerator Statement: Patient Status (alive or Deceased) at follow-up AND neurologic status with an assessment using the NIH Stroke Scale (by an examiner who is certified by the American Stroke Association)
Denominator Statement: CARE Registry patients that underwent carotid endarterectomy
Denominator Exclusions: Patients with a discharge status of deceased.
Patients with was an acute, evolving stroke and dissection during the episode of care.

Measure Type: Process
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Facility, Population : National

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a *process or intermediate outcome* measure is that it is based on a systematic review (SR) and grading of the body of empirical evidence where the specific focus of the evidence matches what is being measured.

The developer provides the following evidence for this measure:

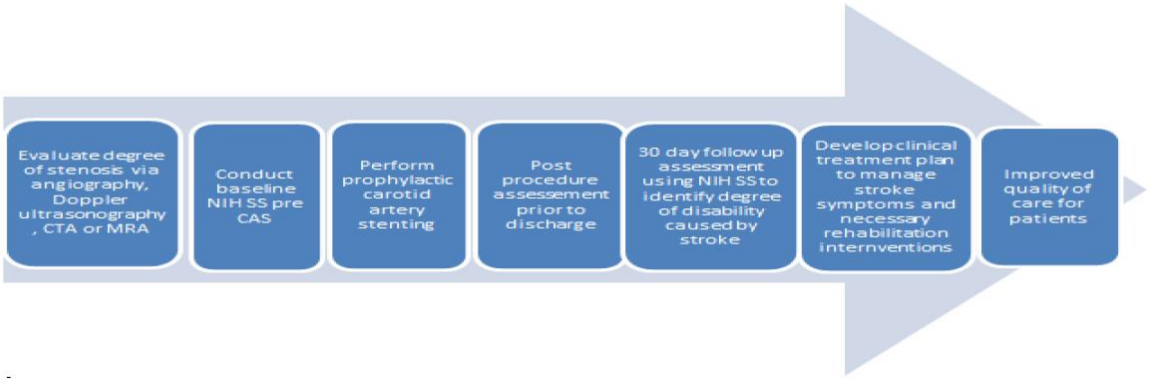
- **Systematic Review of the evidence specific to this measure?** ☒ **Yes** ☐ **No**
- **Quality, Quantity and Consistency of evidence provided?** ☒ **Yes** ☐ **No**
- **Evidence graded?** ☒ **Yes** ☐ **No**

Evidence Summary

- This is facility- and population-level measure calculates proportion of patients with carotid endarterectomy procedures who had follow up performed for evaluation of vital status and

neurological assessment with an NIH Stroke Scale (by an examiner who is certified by the American Stroke Association).

- The developer provided the following path to demonstrate the path between the process and health outcome of improved quality of care:



- The rationale for this measure is supported by two clinical guideline recommendations:
  1. SCAI/SVMB/SVS: *"Monitoring of outcomes with independent post-procedural neurological assessment using standardized instruments and definitions is critically important to ensure high-quality intervention and patient safety. Institutions offering carotid stent placement must have a quality assurance program specifically designed to assess the results of carotid interventions in their locale. The integrity and accuracy of outcome reporting is reliant on the incorporation of mandatory independent and objective neurologic assessment by a qualified and NIH Stroke Scale-certified individual for all patients undergoing carotid stenting".* No grade assigned.
  2. ASA/ACCF/AHA/AANN/AANS/ACR/ASNR/CNS/SAIP/SCAI/SIR/SNIS/SVM/SVS: *"Noninvasive imaging of the extracranial carotid arteries is reasonable 1 month, 6 months, and annually after revascularization to assess patency and exclude the development of new or contralateral lesions. Once stability has been established over an extended period, surveillance at extended intervals may be appropriate. Termination of surveillance is reasonable when the patient is no longer a candidate for intervention".* Grade: Level C- weight of evidence/opinion is in favor of usefulness/efficacy. It is reasonable to perform procedure. The developer provides the Quantity/Quality/Consistency for the imaging clinical guideline.
- As an additional source of evidence to support the measure, the developer provided information from a Statement from for Healthcare Professionals From the American Heart Association/American Stroke Association and endorsed by the Society of Vascular and Interventional Neurology.
- The developer performed an Up-to-date literature review and additional sources of evidence were provided.

Exception to evidence

NQF guidance provides that when there is insufficient empirical evidence in support of a measure, a determination should be made about whether there are, or could be, performance measures of a related outcome or evidence-based intermediate clinical outcome or process. Currently, there are no NQF-endorsed measures that assess the proportion of patients with carotid endarterectomy procedures who had follow up performed for evaluation of vital status and neurological assessment with an NIH Stroke Scale. If the Committee determines that there is an appropriate alternate measure, an exception to the evidence would not be warranted. Alternatively, if the Committee does not identify an appropriate alternate measure, it may agree that it is OK (beneficial) to hold providers accountable for performance in the

absence of empirical evidence of benefits to patients, in which case it would rate the evidence as insufficient with exception.

Guidance from the Evidence Algorithm
Process measure with mostly tangential evidence (Box 3) → Evidence not graded (Box 7) → A measure of a related outcome may not exist (Box 10) → Systematic assessment of expert opinion (Box 11) → If Committee agrees it is OK/beneficial to hold providers accountable for performance in the absence of empirical evidence of benefits to patients → rate as INSUFFICIENT WITH EXCEPTION

Questions for the Committee:

- *What is the relationship of this measure to patient outcomes?*
- *How strong is the evidence for this relationship?*
- *Is the evidence directly applicable to the process of care being measured, including specific use of the NIH Stroke Scale?*

○ *For possible exception to the evidence criterion:*

- *Are there, or could there be, performance measures of a related health outcome, OR evidence-based intermediate clinical outcomes, intervention/treatment?*
- *Is there evidence of a systematic assessment of expert opinion beyond those involved in developing the measure?*

- *Does the SC agree that it is acceptable (or beneficial) to hold providers accountable without empirical evidence?*

Preliminary rating for evidence: ☐ High ☐ Moderate ☐ Low ☒ Insufficient

1b. Gap in Care/Opportunity for Improvement and 1b. Performance Gap

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- Performance scores are not provided on the measure as specified.
- Data on use of the stroke scale is available in the testing section of this measure submission. 12, 447 patients were included in the testing sample.

Disparities

- The developer did not provide any data on disparities, but noted that literature has shown that all races may not have opportunity for equal care.

Questions for the Committee:

○ *Is there a gap in care that warrants a national performance measure?*

○ *If no disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- Process measure; not clear if it actually assesses the outcome.
- Why exclude in-hospital deaths when they are accounted for in the numerator?
- Facility level only
- Sponsored by ACC
- How does this harmonize with the SVS measure?
- I have concerns about the overall measure construct as it is currently specified and tested.
- Good outcome measures based on patient reported outcome tools or clinician reported assessment tools can be built and implemented into clinical practice. Along with these measures it is a good idea to have an accompanying paired process measure to understand the rate at which the assessments are being performed. The process of simply administering a tool really can't stand by itself and should be considered with an outcome measure.
- The evidence stated is insufficient. The first citation relates to an ungraded general guideline recommendation to monitor neurological outcomes and the second relates to non-invasive imaging which is not a part of this measure. If this was an outcome measure, then some of the steps in the process flow make sense, but it is simply a process of administering a tool and assessing for mortality which are not linked strongly to the desired outcome. The outcome itself is not specific "Improved Quality of Care". How do you measure and know that you've reached that outcome. Measure would be so much better/ stronger if was using the NIH stroke scale to actually measure an outcome within 30 or 60 days post discharge.
- I don't think that the measure as currently specified is of strong value to endorse with an insufficient with exception status.
- This is a PROCESS measure of patients who underwent carotid endarterectomy who had an evaluation of the NIH stroke Scale by a certified examiner after the procedure.
- This is a facility/population level measure.
- To participate in this measure, a hospital would need to participate in the NCDR CARE registry. I would request more information about the number of hospitals that participate in this registry and the reliability of 30-day follow-up.
- While guidelines from both SCAI/SVMB/SVS and ASA/ACCF/AHA/AANN/AANS/ACR/ ASNR?CNS/SAIP/SCAI/SIR/SNIS/SVM/SVS, state that this standardize examination is critical to determine outcomes for the CEA procedure, they do not report that there is data showing that this PROCESS (undergoing NIH Stroke Exam) leads to better outcomes than post-CEA patients that do not undergoing this examination.
- Class IIa: weight of evidence/opinion is in favor of usefulness/efficacy.
- Level of Evidence: C: recommendation was consensus opinion, case studies, or standard of care

1b.

- Question the overall value of the process measure construct, particularly the recording of alive or deceased. Is the goal really to understand the mortality rate following this procedure? Again, recommend having a distinct, definable timeframe for follow-up with the patient.
- Although the data presented is not entirely representing the measure as specified, it appears that there is opportunity for improving the follow-up rates with patients and the use of the NIH stroke tool; less than 2% of the patients had this NIHSS tool recorded."
- "Performance Gap information is not included in the performance gap section of this measure application.
- In testing information provided on page 29, the developers report that of 12,447 patients in their registry, 242 (1.94%) had a post-procedure NIHSS recorded within 30 days after CEA. 30-day follow up was 58.17%. Preoperative NIHSS was missing in 11,295/12,447 patients (90.7%)

1c.

- Although this measure was not labeled as a composite, there are two components required to meet the numerator. The recording of patient status as alive or deceased and the administration of an assessment tool, NIHSS.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s): NCDR Care Registry

Specifications:

- This is a facility- and population-level measure that calculates proportion of patients with carotid endarterectomy procedures who had follow up performed for evaluation of vital status and neurological assessment with an NIH Stroke Scale. The care setting is hospital. Data elements appear to be clearly defined in the measure information form and the NCDR CARE registry data dictionary is provided. There is no risk adjustment or stratification.

Questions for the Committee :.

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Is the logic or calculation algorithm clear?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- The signal-to-noise analysis, which is appropriate for this type of measure, differentiates the true difference between measured entities (the signal) to random measurement error (the noise). A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in between hospital performance. A value of 0.7 is often regarded as a minimum acceptable reliability value. The table demonstrates very high reliability and is provided based on CAS procedural volumes, with reliability increasing in higher volumes.
- The testing sample included a cohort of the NCDR CARE Registry (12,477 patients).

- The developers used a test-retest (inter-rater reliability) to assess the accuracy of the data collection methodology and applicable patient characteristic data elements entered into the CARE registry including age, gender, race, smoking, history of PAD, diabetes, chronic lung disease & dyslipidemia. To construct the test/retest, developers identified 449 patient during the testing period with 2 procedures in facilities completing >30 procedures to produce more reliable estimates of the data elements.
- Data element definitions and differences are provided for each data element. Age, gender and race did not vary between the test/retest populations, and smoking, history of PAD, diabetes, chronic lung disease &

dyslipidemia variation were all < 2.7%. Note: NQF prefers to see kappa values in addition to the percent agreement provided.
- The measure is specified such that cases with missing data are assumed to have not met the metric.

Results of reliability testing
For the elements assessed, the results reported in the NCDR CARE record appear to be consistent and accurate, though it is unclear which data set was used for the signal to noise analysis.

Level	Signal-to-Noise
All, >10 Procedures	.982
>Q1 (>77 Procedures)	.985
>Q2 (>166 Procedures)	.994
>Q3 (>315 Procedures)	.995
>Average (>234 Procedures)	.996

Guidance from the Reliability Algorithm:
Precise specs (Box 1) → empiric reliability testing (Box 2) →testing of measure score (Box 4)→appropriate method of testing (Box 5) →testing results (Box 6) → high certainty of reliability (Box 6a) → HIGH

Questions for the Committee:
- *Is the test sample adequate to generalize for widespread implementation?*
- *Does limiting testing to > 30 procedures per facility impact the measure reliability?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Preliminary rating for reliability: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ Yes ☒ Somewhat ☐ No

- The measure specifications allow for follow-up and NIHSS reassessment between 21-60 days inclusive, and the evidence states 30 days.

Question for the Committee:
- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

SUMMARY OF TESTING
Validity testing level ☒ Measure score ☐ Data element testing against a gold standard ☐

Both

Method of validity testing of the measure score:

- ☒ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method:

- The developer described face validity by expert cardiologist panel who actively perform CAS and carotid endartectomy procedures including leading experts in the field and vetted the measure with 3 committees and the 16 member NCDR Management Board and 31 member ACCF Board of Trustees prior to NQF submission. Face validity of a performance measure—the subjective determination by experts that, on the face of it, the measure appears to reflect quality of care is the weakest demonstration of validity. The experts assessing face validity must agree that the **computed measure score** from the measure as specified can be used to distinguish good and poor quality. Based on the testing information submitted, the measure does not meet NQF requirements for face validity.
- The developer described content validity by stating that numerous studies assess the comparative effectiveness of stenting vs. surgery for carotid occlusion on the outcomes of myocardial infarction, death and stroke. The measure, as specified, looks for information about evaluation of vital status and neurologic assessment with an NIH Stroke Scale (by an examiner who is certified by the American Stroke Association). Based on the information provided and the measure focus, this is not considered to be sufficient validity testing of the measure.
- No formal statistical validity tests were provided.

Validity testing results:

As this measure is being proposed primarily on the basis of its content validity, as described above, there are no empiric results from formal validity testing.

Questions for the Committee:

- *Do the face and content validity provided demonstrate the measure has sufficient validity so that conclusions about quality can be made?*
- *Do you agree that the score from this measure as specified is an indicator of quality?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

The only proposed exclusion is for patients being treated in the context of an acute evolving stroke. The developers provide data to support that the measure that does not suffer from excluding those with an acute ischemic stroke.

Questions for the Committee:

- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☒ **None** ☐ **Statistical model** ☐ **Stratification**

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- The developer notes an association between procedural volume and follow up rates; higher volume centers perform substantially better than poor performing centers. The developer also states that performance rates varied from 0-50% therefore it is feasible for hospitals to assess all of their patients.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

- This is not applicable as there is only one data source/specification for this measure.

2b7. Missing Data

- The developer describes the NCDR Data Quality Report that assesses for data completeness, consistency and accuracy (integrity), scoring provided data with 3 "light" levels: Red lights (submission failure for integrity and completeness check – data not processed), yellow lights (passed for integrity and fails completeness – requires data resubmission), and green lights (passed all quality checks and included for aggregate computations). No sampling of NCDR patient data is allowed as registry inclusion mandates 100% consecutive patients.
- The developer indicates that the measure is specified such that cases with missing data are assumed to have not met the metric.

Guidance from the Validity Algorithm:

Consistent with specifications (Box 1) →potential threats to validity addressed (Box 2) →face validity possibly assessed, but information regarding method and results not provided (Box 4) → Insufficient

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.

- Threat to reliability. The developer lists the planned registry fields intended for use, not having a clearly stated time for follow-up. Field specs state "recommended timeframe for follow-up is 30 days, but later states in submission document that system allows 21 to 60 days. Specs should clearly state (and apply) a set definition of the window.
- The field follow-up performed does not indicate what type of follow-up occurred (dead or alive assessment or the administration of the NIH Stroke scale). Can't assume both were completed, yet both are part of the numerator statement. Later there is a status field (alive or deceased) but no date for the assessment of alive, only a date of death.
- Additional treat to reliability is the exclusion for evolving stroke … how is this defined and consistently removed from the denominator?
- I do not understand how the registry tracks 30-day follow up and how this follow-up is reliable.
- I do not understand how the registry verifies that an examiner is certified with the NIHSS
- Can the developers please provide this information?

2a2.

- Could the developer further explain the method for reliability score testing that was used? As a review I don't know what "all 10 procedures" or Q1, Q2, Q3 or the average represent.
- "Reliability was determined by testing measure score.
- Reliability testing of clinical and demographic data elements was provided with percentages (no kappa values).
- Reliability testing of 30-day follow up and of the PROCESS in question, performance of NIHSS exam was not provided.
- Reliability testing was not provided on HOW the registry verifies that an examiner is NIHSS certified

2b.

- Note that follow-up of vital status is an assessment of alive or deceased. Also that the timeframe for follow-up could be stated clearly. The registry algorithm credits 21 to 60 days, the recommendation is follow-up within 30 days.
- Follow-up as a field within the registry listed for measure calculation is vague, does not indicate what type of follow-up occurred. Another field not listed in the measure calculation (seq 9012 the date of the NIH tool) would be more accurate in calculating the process of administering a tool within a specified timeframe.

2b2.

- Data element validity testing result supplied did not at all relate to any of the data elements used to calculate this measure. Developer reviewed age, gender, race, smoking, PAD, diabetes, chronic lung disease and dyslipidemia, none of which are used for measure calculation.
- Face validity only.

2b3.

- Insufficient reliability and validity testing. Threats to measure include lack of testing, validation of critical data elements, specifications for numerator and exclusions.

Criterion 3. Feasibility

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The data source used to collect and calculate measure performance is the NCDR Care Registry. The developer states that all data elements are in defined fields in electronic clinical data, and may be collected via third-party vendors. The specifications are available in the public domain.
- The developer indicated that the 2016 the annual pricing for hospitals, NCDR Analytic and Reporting Services, and licensing of measure specifications is $4,530.00.
- The majority of the required data elements are routinely generated and acquired during the delivery of care to this patient population. Electronic extraction of data recorded as part of the procedure expedites data collection. The data elements required for this measure are readily available within the patient's medical record or can be attained without undue burden within the hospital. Most data elements exist in a structured format within patient's electronic health record.
- A similar measure (NQF#1531) was reviewed in 2010. The committee reviewing the measure had concerns about feasibility and the burden of data collection on organizations for capturing the assessment at the follow-up visits since the assessment does not occur during hospitalization. There was also concern that the assessment must be performed by an examiner who is certified by the American Stroke Association

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Are the required data elements available in electronic form, e.g., EHR or other electronic sources?*

Is the data collection strategy ready to be put into operational use?

Preliminary rating for feasibility: ☐ **High** ☐ **Moderate** ☒ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

- Feasibility concerns related to current technical accuracy of measure as specified and overall value in terms of 1) need for two components and 2) process measure for tool administration alone. Good outcome measures based on patient reported outcome tools or clinician reported assessment tools

can be built and implemented into clinical practice. Along with these measures it is a good idea to have an accompanying paired process measure to understand the rate at which the assessments are being performed. The process of simply administering a tool really can't stand by itself and should be considered with an outcome measure.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☒ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details

- The measure is not currently publicly reported but the developers state that the plan is for this measure to be publicly reported in the future. The developer provides a plan for future use/ implementation and is applying to be a CMS Qualified Entity to allow developer's access to CMS reporting. Further clarification is needed.

Improvement results Not available

Unexpected findings (positive or negative) Not available

Potential harms None identified

Feedback :

None identified

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

See comments in feasibility section.

Criterion 5: Related and Competing Measures

Related measure

This is specified identically to NQF# 2396 Carotid artery stenting: Evaluation of Vital Status and NIH Stroke Scale at Follow Up, with the exception of the CAS versus the CEA population.

Harmonization

The measures are harmonized.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number

Measure Title: Carotid Endarterectomy: Evaluation of Vital Status and NIH Stroke Scale at Follow Up

IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 5/31/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the

measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☐ Health outcome: Click here to name the health outcome

☐Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☒ Process: Carotid Endarterectomy: Evaluation of Vital Status and NIH Stroke Scale at Follow Up

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

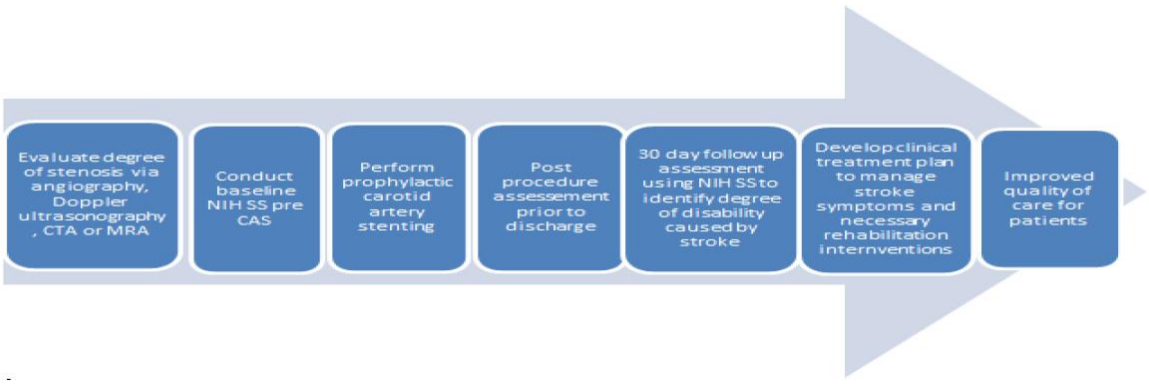
1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

_____

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.



1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☒ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☒ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

#1) Rosenfield K, Cowley MJ, Jaff MR, Ouriel K, Gray W, Cates CU, Feldman T, Babb JD, Gallagher A, Green R, Kent KC, Roubin GS, Weiner BH, White CW. SCAI/SVMB/SVS clinical competence statement on carotid stenting: training and credentialing for carotid stenting— multispecialty consensus recommendations, a report of the SCAI/SVMB/SVS writing committee to develop a clinical competence statement on carotid interventions. J Am Coll Cardiol 2005;45:165–74.

URL for Clinical Competence Statement: http://content.onlinejacc.org/article.aspx?articleid=1136222

#2) Brott TG, Halperin JL, Abbara S, et al. 2011

ASA/ACCF/AHA/AANN/AANS/ACR/ASNR/CNS/SAIP/SCAI/SIR/SNIS/SVM/SVS Guideline on the Management of Patients With Extracranial Carotid and Vertebral Artery Disease: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American Stroke Association, American Association of Neuroscience Nurses, American Association of Neurological Surgeons, American College of Radiology, American Society of Neuroradiology, Congress of Neurological Surgeons, Society of Atherosclerosis Imaging and Prevention, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of NeuroInterventional Surgery, Society for Vascular Medicine, and Society for Vascular Surgery Developed in Collaboration With the American Academy of Neurology and Society of Cardiovascular Computed Tomography. J Am Coll Cardiol. 2011;57(8):e16-e94.URL for guideline:

http://content.onlinejacc.org/article.aspx?articleid=1144187

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

#1) Page 9 of 10 “Monitoring of outcomes with independent post-procedural neurological assessment using standardized instruments and definitions is critically important to ensure high-quality intervention and patient safety. Institutions offering carotid stent placement must have a quality assurance program specifically designed to assess the results of carotid interventions in their locale. The integrity and accuracy of outcome reporting is reliant on the incorporation of mandatory independent and objective neurologic assessment by a qualified and NIH Stroke Scale-certified individual for all patients undergoing carotid stenting”.

#2) Page 40 of 79, e55 “Noninvasive imaging of the extracranial carotid arteries is reasonable 1 month, 6 months, and annually after revascularization to assess patency and exclude the development of new or contralateral lesions. Once stability has been established over an extended period, surveillance at extended intervals may be appropriate. Termination of surveillance is reasonable when the patient is no longer a candidate for intervention”.

1a.4.3. Grade assigned to the quoted recommendation <u>with definition</u> of the grade:

The consistent assessment of outcomes using a standardized approach concordant with that used in trials is fundamental to generating meaningful outcomes performance benchmarks and comparisons with the trial outcomes. Regardless of specific guideline recommendations, this measure is necessary as the foundation of future outcomes metrics.

#1) No grade assigned

#2) Class IIa,(Definition of Class IIa: Weight of evidence/opinion is in favor of usefulness/ efficacy. IT IS REASONABLE to perform procedure/administer treatment.

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

#2) See table below.

ESTIMATE OF CERTAINTY (PRECISION) OF TREATMENT EFFECT	CLASS I *Benefit >>> Risk* Procedure/Treatment **SHOULD** be performed/ administered	CLASS IIa *Benefit >> Risk* *Additional studies with focused objectives needed* **IT IS REASONABLE** to perform procedure/administer treatment	CLASS IIb *Benefit ≥ Risk* *Additional studies with broad objectives needed; additional registry data would be helpful* Procedure/Treatment **MAY BE CONSIDERED**	CLASS III *No Benefit* or CLASS III *Harm* Procedure/ Test / Treatment COR III: No benefit — Not Helpful — No Proven Benefit COR III: Harm — Excess Cost w/o Benefit or Harmful — Harmful to Patients
LEVEL A Multiple populations evaluated* Data derived from multiple randomized clinical trials or meta-analyses	■ Recommendation that procedure or treatment is useful/effective ■ Sufficient evidence from multiple randomized trials or meta-analyses	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from multiple randomized trials or meta-analyses	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Sufficient evidence from multiple randomized trials or meta-analyses
LEVEL B Limited populations evaluated* Data derived from a single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is useful/effective ■ Evidence from single randomized trial or nonrandomized studies	■ Recommendation in favor of treatment or procedure being useful/effective ■ Some conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation's usefulness/efficacy less well established ■ Greater conflicting evidence from single randomized trial or nonrandomized studies	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Evidence from single randomized trial or nonrandomized studies
LEVEL C Very limited populations evaluated* Only consensus opinion of experts, case studies, or standard of care	■ Recommendation that procedure or treatment is useful/effective ■ Only expert opinion, case studies, or standard of care	■ Recommendation in favor of treatment or procedure being useful/effective ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation's usefulness/efficacy less well established ■ Only diverging expert opinion, case studies, or standard of care	■ Recommendation that procedure or treatment is not useful/effective and may be harmful ■ Only expert opinion, case studies, or standard of care
Suggested phrases for writing recommendations	should is recommended is indicated is useful/effective/beneficial	is reasonable can be useful/effective/beneficial is probably recommended or indicated	may/might be considered may/might be reasonable usefulness/effectiveness is unknown/unclear/uncertain or not well established	COR III: No Benefit is not recommended is not indicated should not be performed/ administered/ other is not useful/ beneficial/ effective COR III: Harm potentially harmful causes harm associated with excess morbidity/mortality should not be performed/ administered/ other
Comparative effectiveness phrases†	treatment/strategy A is recommended/indicated in preference to treatment B treatment A should be chosen over treatment B	treatment/strategy A is probably recommended/indicated in preference to treatment B it is reasonable to choose treatment A over treatment B		

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**
For Recommendation #2) ACCF/AHA Task Force on Practice Guidelines. Methodology Manual and Policies From the ACCF/AHA Task Force on Practice Guidelines. American College of Cardiology Foundation and American Heart Association, Inc. Cardiosource.com. 2010. Available at: http://assets.cardiosource.com/Methodology_Manual_for_ACC_AHA_Writing_Committees.pdf and http://my.americanheart.org/idc/groups/ahamah-public/@wcm/@sop/documents/downloadable/ucm_319826.pdf

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☒ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

The evidence review focused on the importance of 30 day follow up assessment for this patient population to determine morbidity in terms of incidence and prevalence of stroke and mortality that may be associated with carotid artery revascularization via stenting.

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

Level of Evidence: C (Definition of Evidence Level C: primary source of the recommendation was consensus opinion, case studies, or standard of care).

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

Guideline #2) listed in section 1.a. 4.2
Evidence Level A: data were derived from multiple randomized clinical trials or meta-analyses.
Evidence Level B: data were derived from a single randomized trial or nonrandomized studies.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: 2001 - 2006

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

One multinational, prospective, randomized study of 1,214 patients.

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

Certainty or confidence in the estimates (verbatim from Eckstein et al.,cited below)

In both the intention-to-treat and per-protocol analyses the Kaplan-Meier estimates of ipsilateral ischemic strokes up to 2 years after the procedure and any periprocedural stroke or death do not differ between the carotid artery stenting and the carotid endarterectomy groups (intention to treat 9.5%vs 8.8%; hazard ratio (HR) 1.10, 95%CI 0.75 to 1.61; log-rank p=0.62; per protocol 9.4%vs 7.8%; HR 1.23, 95%CI 0.82 to 1.83; log-rank p=0.31).

In both the intention-to-treat and per-protocol populations, recurrent stenosis of 70% or more is significantly more frequent in the carotid artery stenting group compared with the carotid endarterectomy group, with a life-table estimate of 10.7% versus 4.6% (p=0.0009) and 11.1% versus 4.6% (p=0.0007), respectively.

Eckstein H.H., Ringleb P., Allenberg J.R.; Results of the Stent-Protected Angioplasty versus Carotid Endarterectomy (SPACE) study to treat symptomatic stenosis at 2 years: a multinational, prospective, randomized trial. *Lancet Neurol*. 7 2008:893-902.

Indirectness of studies to the measure focus

The guidelines cited in 1a.4.1. stress the importance of follow up to be conducted upon this patient population. While implied, it can be considered a fundamental aspect of this follow up process to determine the vital status of a patient for this follow up visit. In addition to vital status, this measure requires the NIH Stroke Scale to be performed during the follow up visit, an implied component of follow-up for morbidity.

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

Only two incidences of recurrent stenoses after carotid artery stenting led to neurological symptom. After 2 years' follow-up, the rate of recurrent ipsilateral ischaemic strokes reported in the SPACE trial is similar for both treatment groups

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

One outcome identified by this study is that the incidence of recurrent carotid stenosis at 2 years (identified by ultrasound), was significantly higher after carotid artery stenting then when a CEA was performed. This reinforces the importance of continued follow up on this patient population.

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

In a Statement for Healthcare Professionals From the American Heart Association/American Stroke Association and endorsed by the Society of Vascular and Interventional Neurology, several metrics were proposed intended to provide a framework for standardized data collection at comprehensive stroke centers (CSCs) to facilitate local quality improvement efforts and to allow for analysis of pooled data from different CSCs that may lead to development of national performance standards for CSCs in the future.

1a.8.1 What process was used to identify the evidence?

A literature search was performed of the medical database www.UpToDate.com and the American Heart Association/American Stroke Associate Stroke [http://stroke.ahajournals.org] webpage using keywords: “Carotid Artery Stenosis” and “Follow Up”.

1a.8.2. Provide the citation and summary for each piece of evidence.

The following is quoted verbatim from the reference cited.

Reference: Leifer D, Bravata DM, Connors JJ 3rd, Hinchey JA, Jauch EC, Johnston SC, Latchaw R, Likosky W, Ogilvy C, Qureshi AI, Summers D, Sung GY, Williams LS, Zorowitz R; on behalf of the American Heart Association Special Writing Group of the Stroke Council, Atherosclerotic Peripheral Vascular Disease Working Group, Council on Cardiovascular Surgery and Anesthesia, and Council on Cardiovascular

Nursing. Metrics for measuring quality of care in comprehensive stroke centers: detailed follow-up to Brain Attack Coalition comprehensive stroke center recommendations: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2011;42: Retrieved from http://stroke.ahajournals.org/content/early/2011/01/13/STR.0b013e318208eb99.full.pdf+html on December 4, 2013

Metric 10

Percentage of patients undergoing carotid endarterectomy (CEA), or carotid angioplasty or stenting, with stroke or death within 30 days of the procedure.

Numerator: Number of patients who have a stroke or die within 30 days of CEA, or who have carotid angioplasty or stenting performed because of atherosclerotic disease.

Denominator: Total number of patients who undergo CEA or who undergo carotid angioplasty or stenting because of atherosclerotic disease.

The metric should be calculated for all procedures taken together and separately for the following groups of patients:
(1) symptomatic patients undergoing endarterectomy;
(2) symptomatic patients undergoing carotid angioplasty or stenting; (3) asymptomatic patients undergoing endarterectomy; and
(4) asymptomatic patients undergoing carotid angioplasty or stenting.

Strokes should be included if they meet the clinical definition of a focal neurological deficit that persists for > or equal to 24 hours without other cause or if there is a focal deficit that lasts for a shorter period of time but is associated with an appropriately located acute ischemic lesion on MRI. Clinically silent acute lesions detected on diffusion-weighted MRI should not be included as complications, because they are likely to be common when MRI is performed, although their incidence and clinical significance are uncertain. Patients with confusion or encephalopathy who have multiple punctate lesions that together may explain their clinical findings should also be included as having had a stroke. Published clinical trials about complications after carotid procedures and other interventions have typically used clinical stroke as the end point and other ongoing trials also are using clinical end points. This definition of stroke will apply to this metric and subsequent ones.

This metric is limited to patients with atherosclerotic disease to ensure that the metric encompasses a uniform population of patients.

Justification: The AHA/ASA guidelines for patients with recent TIA or ischemic stroke within the past 6 months and ipsilateral severe (70% to 99%) carotid artery stenosis recommend endarterectomy by a surgeon with a perioperative morbidity and mortality rate of <6% ***(Class I; Level of Evidence A)***. For patients with recent TIA or ischemic stroke and ipsilateral moderate (50% to 69%) carotid stenosis, CEA is recommended, depending on patient-specific factors such as age, sex, comorbidities, and severity of initial symptoms if the perioperative morbidity and mortality risk is estimated to be <6% ***(Class I; Level of Evidence B)***.

Among patients with symptomatic severe stenosis (>70%) in whom either the stenosis is difficult to access surgically, medical conditions are present that greatly increase the risk for surgery, or other specific circumstances exist such as radiation-induced stenosis or restenosis after CEA, the use of carotid angioplasty and stent placement is not inferior to endarterectomy and may be considered ***(Class IIb;***

Level of Evidence B). The procedure is reasonable when performed by operators with established periprocedural morbidity and mortality rates of 4% to 6% ***(Class IIa; Level of Evidence B)***.

The role of carotid angioplasty/stenting in asymptomatic patients has not been established. The AHA/ASA "Guidelines for the Primary Prevention of Stroke" state, "The usefulness of CAS [carotid angioplasty/stenting] as an alternative to CEA in asymptomatic patients at high risk for the surgical procedure is uncertain ***(Class IIb; Level of Evidence C)***." In this setting, if centers choose to perform carotid angioplasty/stenting on asymptomatic patients, the 30-day rate of stroke and death should be tracked separately for such patients and monitored carefully.

The recommended end point to be ascertained after carotid angioplasty and stent placement is any stroke or death within 30 days, to remain consistent with the data collected for CEA. This end point has been used in trials of carotid angioplasty and stenting. For comparable patients, the complication rate for stenting should be similar to that for endarterectomy if stenting is to be a reasonable option. In particular, the complication rate should be expected to be between 4% and 6% for symptomatic >70% stenosis. If carotid angioplasty and stenting are performed, therefore, careful attention must be paid to complication rates, so it is important for CSCs to monitor these rates.

The risk of stroke and death after carotid revascularization are important and can substantially influence the net benefit of the procedure. Assessment and reporting of the "outcome" of stroke for carotid revascularization procedures is not consistent in the absence of a clinical assessment using a standardized stroke scale, or by using claims data. A class IIa, LOE: C guideline advises noninvasive imaging of the extracranial carotid arteries be performed at 1 month, 6 months, and annually after revascularization to assess patency and exclude the development of new or contralateral lesions, it can be implied that patients will have a clinic/office follow-up visits as a follow-up to revascularization procedures. This office visit provides the opportunity for appropriate clinical assessment for key revascularization endpoints, including stroke or death. A process measure that uses a standard assessment of neurologic function, by an examiner who is certified by the American Stroke Association, is a measure that provides feedback on the ability to clearly and accurately assess for, capture and report the incidence of stroke after carotid revascularization procedures. The NIHSS is both reliable and valid, and has become a standard stroke impairment scale for use in both clinical trials and as part of clinical care in the United States [6].

When centers that perform carotid revascularization properly assess patients for adverse events (particularly for stroke) after carotid revascularization, they trigger further evaluation, if necessary. If the 30 day NIH stroke scale is (1) changed from baseline; or (2) abnormal in absence of a baseline, pre-procedure exam, then there should be some documentation on whether or not the abnormal stroke scale represents a new clinical neurological event, and should result in an evaluation by a neurologist.

According to the CARE Registry institutional outcomes reports, the median length of stay for CAS and CEA procedures is one day. This short hospital stay reflects difficulty in reporting "in-hospital" stroke outcomes as a relevant measure. Following carotid artery stenting, patients are typically discharged in one to two days. In a study that evaluated the timing of complications following CAS, 53% of postoperative events/complications occurred within 6 hours of CAS, 5.3% between 6 and 12 hours, 8% between 12 and 24 hours, and 34.2% >24 hours post procedure [1]. Late events >24 hours were access-site-related and neurologic events.

The primary endpoints of major contemporary trials used 30 day events (stroke, MI* or death) and included neurologic evaluation to identify stroke. Based on trial endpoints, 30 day outcomes have greater importance. Post-procedure stroke is one of the major adverse outcomes from carotid artery stenting and carotid endarterectomy. For example, this was the major outcome in the recent CREST trial, a randomized comparison of carotid artery stenting and surgical endarterectomy. A recent meta-analysis by Murad et al. summarized the results of 13 randomized controlled trials to assess the comparative effectiveness of stenting vs. surgery for carotid occlusion on the outcomes of myocardial infarction, death and stroke [13]. The latter of these outcomes are proposed to serve as a process measure for quantifying the quality of carotid revascularization by this measure.

There is a sound clinical rationale for systematically measuring the outcomes of carotid revascularization. First, without knowing the outcomes, a hospital cannot know if it is applying its treatment in a safe and effective manner. Given how infrequently current providers assess the 30-day survival and stroke outcomes, it is obvious that more than half of these hospitals have no foundation with which to assess the quality of their care. We have proposed a process measure, merely assessing the stroke-free survival of treated patients, because without more clear ascertainment of outcomes it is not possible to provide risk-adjusted comparisons across centers and provide clear benchmarks of performance to identify hospitals that have the opportunity to improve. Second, as the country seeks to support the use of evidence-based medicine, the majority of the evidence in carotid disease comes from clinical trials. However, many of the trials establishing the benefits of carotid revascularization require that centers document a certain success rate, without complications of stroke or death, before the center can participate in a clinical trial. If a center does not know its rate, it will not know whether or not the benefits observed in a clinical trial apply to their practice. Finally, for patients to be adequately informed about the risks and benefits of treatment, hospitals need to have reliable data to share with their patients. By collecting, analyzing and reporting the outcomes of treatment, hospitals will be much better able to provide their patients the information that they need to make a treatment decision.

Stroke is the second leading cause of all hospital admissions among older patients and the leading reason for neurology-related admissions. From 1999 to 2009, the number of inpatient discharges from short stay hospitals with stroke as the first-listed diagnosis has remained stable with 961,000 discharges in 1999 and 971,000 discharges in 2009 (National Heart, Lung, and Blood Institute [NHLBI] tabulation, National Hospital Discharge Survey [NHDS], National Center for Health Statistics [NCHS]). Correspondingly, stroke death rates fell by 24% from 1994 to 2004. This decline suggests that there have been general improvements in the management of patients with acute stroke, decreases in the severity of stroke and/or improved detection or coding of milder stroke cases. Part of the decline in hospital stroke mortality may be due to the shorter length of stay resulting in more out of hospital death. The greatest risk of mortality for patients with stroke occurs in the first 30 days, with case-fatality rates ranging from 8% to 20% for ischemic stroke, with substantially higher rates for stroke due to subarachnoid or intracerebral hemorrhage (as high as 50%). The immediate cause of death in more than
60% of stroke cases is thought to be related to complications of the stroke itself. After the first week, cardiac causes, pneumonia, pulmonary embolism, sepsis, and other medical complications account for the majority of the stroke-related mortality. In 2008, approximately 46% of all stroke deaths occurred in the hospital (unpublished NHLBI tabulation of NCHS 2008 Mortality Data Set). The annual U.S. economic burden of stroke is estimated at $20.4 billion for direct and $53.6 billion indirect costs. [7]

1] Timing and frequency of complications after carotid artery stenting: what is the optimal period of observation? Tan KT, Cleveland TJ, Berczi V, McKevitt FM, Venables GS, Gaines PA. J Vasc Surg.
2003;38(2):236

2] 30 day results from the SPACE trial of stent-protected angioplasty versus carotid endarterectomy in symptomatic patients: a randomised non-inferiority trial. SPACE Collaborative Group, Ringleb PA, Allenberg J, Brückmann H, Eckstein HH, Fraedrich G, Hartmann M, Hennerici M, Jansen O, Klein G, Kunze A, Marx P, Niederkorn K, Schmiedt W, Solymosi L, Stingele R, Zeumer H, Hacke W. Lancet.
2006;368(9543):1239
3] Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis. Mas JL,
Chatellier G, Beyssen B, Branchereau A, Moulin T, Becquemin JP, Larrue V, Lièvre M, Leys D, Bonneville JF, Watelet J, Pruvo JP, Albucher JF, Viguier A, Piquet P, Garnier P, Viader F, TouzéE, Giroud M, Hosseini H, Pillet JC, Favrole P, Neau JP, Ducrocq X, EVA-3S Investigators. N Engl J Med. 2006;355(16):1660.
4] Carotid artery stenting compared with endarterectomy in patients with symptomatic carotid stenosis (International Carotid Stenting Study): an interim analysis of a randomised controlled trial. International Carotid Stenting Study investigators, Ederle J, Dobson J, Featherstone RL, Bonati LH, van der Worp HB, de Borst GJ, Lo TH, Gaines P, Dorman PJ, Macdonald S, Lyrer PA, Hendriks JM, McCollum C, Nederkoorn PJ, Brown MM. Lancet. 2010;375(9719):985.
5] Thirty-day outcomes for carotid artery stenting in 6320 patients from 2 prospective, multicenter, high-surgical-risk registries. Gray WA, Chaturvedi S, Verta P, Investigators and the Executive Committees. Circ Cardiovasc Interv. 2009;2(3):159.
6] Goldstein LB, Bushnell CD, Adams RJ, Appel LJ, Braun LT, Chaturvedi S, Creager MA, Culebras A, Eckel RH, Hart RG, Hinchey JA, Howard VJ, Jauch EC, Levine SR, Meschia JF, Moore WS, Nixon JV, Pearson TA; on behalf of the American Heart Association Stroke Council, Council on Cardiovascular Nursing, Council on Epidemiology and Prevention, Council for High Blood Pressure Research, Council on Peripheral Vascular Disease, and Interdisciplinary Council on Quality of Care and Outcomes Research. Guidelines for the primary prevention of stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2011;42:517–584
7] AHA Statistical Update Heart Disease and Stroke Statistics—2012 Update. A Report From the
American Heart Association Circulation. *2012; 125: e2-e220 Published online before print December 15, 2011, doi: 10.1161/CIR.0b013e31823ac046*

8] David C. Costs and cost-effectiveness of carotid stenting vs. endarterectomy for patients at increased surgical risk: Results from the SAPPHIRE trial. Catheter Cardiovasc Interv. 2011; Mar 1;77(4):463-72
9] Mantese VA, Timaran CH, Chiu D, et al. The Carotid Revascularization Endarterectomy versus Stenting
Trial (CREST): stenting versus carotid endarterectomy for carotid disease. Stroke. 2010;41:S31-S34.
10] Mas JL, Trinquart L, Leys D, et al. Endarterectomy Versus Angioplasty in Patients with Symptomatic Severe Carotid Stenosis (EVA-3S) trial: results up to 4 years from a randomised, multicentre trial. Lancet Neurol. 2008;7:885-92.
11] Mast H, Chambless LE, Mohr JP, et al. [Indications for endarterectomy in asymptomatic stenoses of the
internal or common carotid artery--results of the North American ACAS Study]. Zentralbl Chir.
1996;121:1033-5.

12] Ringleb PA, Hacke W. [Stent and surgery for symptomatic carotid stenosis. SPACE study results]. Nervenarzt. 2007;78:1130-7.
13] Murad MH, Shahrour A, Shah N, Montori VM. A systematic review and meta-analysis of randomized trials of carotid endarterectomy vs stenting. J Vasc Surg 2011;53:792-7

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
evidence_attachment_CEA_May_2016_Final.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
There is a sound clinical rationale for systematically measuring the outcomes of carotid revascularization. First, without knowing the outcomes, a hospital cannot know if it is applying its treatment in a safe and effective manner. Given how infrequently current providers assess the 30-day survival and stroke outcomes, it is obvious that more than half of these hospitals have no foundation with which to assess the quality of their care.

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
None

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
the country seeks to support the use of evidence-based medicine, the majority of the evidence in carotid disease comes from clinical trials. However, many of the trials establishing the benefits of carotid revascularization require that centers document a certain success rate, without complications of stroke or death, before the center can participate in a clinical trial. If a center does not know its rate, it will not know whether or not the benefits observed in a clinical trial apply to their practice. Finally, for patients to be adequately informed about the risks and benefits of treatment, hospitals need to have reliable data to share with their patients. By collecting, analyzing and reporting the outcomes of treatment, hospitals will be much better able to provide their patients the information that they need to make a treatment decision.

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
None

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
Literature has shown that all races may not have the opportunity for equal care. Please see specific information in the testing document.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**

Patient/societal consequences of poor quality

1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

Not a PRO_PM measure.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*

Cardiovascular, Neurology : Stroke/Transient Ischemic Attack (TIA)

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

Prevention : Screening

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)

This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*

No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.

Measure not previously endorsed; note that the measure is configured identically to 2396.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*

IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

Patient Status (alive or Deceased) at follow-up AND neurologic status with an assessment using the NIH Stroke Scale (by an examiner who is certified by the American Stroke Association)

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*

One year for numerator and denominator

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
Field Name: Patient Follow-up Performed Seq No: 9000
Definition: Indicate whether patient follow-up was performed after the procedure. The recommended timeframe for follow-up is 30 days; the measure credits any follow up occurring between days 21-60, inclusive.
1=Yes

Field Name: Follow-Up Date Seq No: 9002
Definition: Indicate the date of follow-up. The recommended timeframe for follow-up is 30 days;the measure credits any follow up occurring between days 21-60, inclusive.

Field Name: Follow Up NIH Stroke Scale Administered Seq No: 9010
Definition: Indicate if the National Institutes of Health Stroke Scale (NIHSS) was administered during follow-up occurring between days 21-60, inclusive
1=Yes

Follow-up NIH Stroke Scale Examiner Certified Seq No: 9014

Definition: Indicate the date the National Institutes of Health Stroke Scale (NIHSS) was administered during the follow-up period. Note - The recommended timeframe for follow-up is 30 days; the measure credits any follow up occurring between days 21-60, inclusive.

1=Yes

Field Name: Follow-up NIH Stroke Scale Examiner Certified Seq No: 9014
Definition: Indicate if the examiner who performed follow up is certified to determine the NIH Stroke and is not the operator who performed the current procedure.

Examiner certified= yes
Supporting definitions:
The Stroke Scale assessment should be conducted by someone other than the operator for the current procedure.
Note - NIHSS examiners may become certified through the American Stroke Association.
NIH Stroke Scale Certification is currently available online free of charge: http://learn.heart.org/ihtml/application/student/interface.heart2/nihss.html

Field Name: Patient Status Seq No: 9100
Definition: Indicate if the patient is alive or deceased.

Alive (1) or deceased (2)

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
CARE Registry patients that underwent carotid endarterectomy

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Populations at Risk

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should*

be provided in an Excel or csv file in required format at S.2b)
Count of CARE Registry patients that had a carotid endarterectomy

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Patients with a discharge status of deceased.
Patients with was an acute, evolving stroke and dissection during the episode of care.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
Field Name: Discharge Status Seq No: 8010
Definition: Indicate whether the patient was alive or deceased at discharge from the hospitalization during which the procedure occurred.
Alive=2

Field Name: Spontaneous Carotid Artery Dissection Seq No: 5060
Definition: Indicate if the patient has had a spontaneous carotid artery dissection prior to the current procedure.

1=Yes

Field Name: Acute Evolving Stroke Seq No: 4340
Definition: Indicate if the patient has experienced an acute evolving stroke with ischemia which is ongoing and progressing at the time of the procedure. Acute evolving stroke includes all of the following:
1. Any sudden development of neurological deficits attributable to cerebral ischemia and/or infarction.
2. Onset of symptoms occurring within prior three days and ongoing at time of procedure.
3. The event is marked by progressively worsening symptoms.
Note: Possible symptoms include, but are not limited to the following: numbness or weakness of the face or body; difficulty speaking or understanding; blurred or decreased vision; dizziness; or loss of balance and coordination.

1=Yes

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
The measure is not stratified.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
No risk adjustment or risk stratification
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
No risk adjustment.

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
Not a risk model measure.

S.16. **Type of score:**
Count
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Not a risk model measure.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
Not a PRO=PM measure.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
Not a PRO=PM measure.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Not a PRO=PM measure.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
NCDR Care Registry

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available in attached appendix at A.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility, Population : National

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Not a composite performance measure

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
nqf_testing_attachment_CEA_6_10_Final.docx

NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Carotid Endarterectomy: Evaluation of Vital Status and NIH Stroke Scale at Follow Up
Date of Submission: 5/31/2016
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☐ Outcome (*including PRO-PM*)
☐ Cost/resource	☒ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.
- For information on the most updated guidance on how to address sociodemographic variables and testing in this form refer to the release notes for version 6.6 of the Measure Testing Attachment.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient

preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. 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 sociodemographic 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.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions

15. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (***must be consistent with data sources entered in S.23***)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

We propose to use a clinical registry, the National Cardiovascular Data Registry CARE Registry. This is a national quality improvement registry that is currently participated in by >180 US hospitals. Rigorous quality standards are applied to the data and both quarterly and *ad hoc* performance reports with program-wide benchmarks are generated for participating centers to track and improve their performance.

1.3. What are the dates of the data used in testing? Calendar Year2007Q1-2013Q1

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (***must be consistent with levels entered in item S.26***)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☐ group/practice	☐ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the*

analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample)

A cohort of the NCDR CARE Registry (2007Q1 and 2013 Q1) was used to establish the prevalence of neurological function testing. We restricted our analyses to those hospitals that performed 30 or more carotid revascularization procedures to improve the precision of our estimates. To examine the test/re-stest validity of the measured data elements, we expanded the time window of CARE from 2007-2013 to identify patients with 2 or more procedures.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

A total of 12, 447 patients were included. The process measure that was assessed was vital status and the presence of an NIH Stroke Scale (NIHSS) assessment at 30 days after the CEA procedure. Patients who were acutely experiencing stroke at the time of treatment were excluded, as it is not feasible to distinguish whether or not 30-day neurological outcomes were due to the presenting stroke or due to the treatment. The characteristics of the patients, stratified by collection of their outcomes data, are provided below:

	Total	Both Vital and NIHSS at F/U		
	n = 12447	1 n = 242	0 n = 12205	P-Value
Follow-up Measures				
Alive or deceased recorded	7241 (58.17%)	242 (100.00%)	6999 (57.35%)	< 0.001
NIHSS recorded	242 (1.94%)	242 (100.00%)	0 (0.00%)	< 0.001
A. Demographics				
Age (years, mean +/- standard deviation)	70.78 ± 10.58	69.31 ± 10.41	70.81 ± 10.58	0.029
Sex Male Female Missing	 7375 (59.26%) 5071 (40.74%) 1	 149 (61.57%) 93 (38.43%)	 7226 (59.21%) 4978 (40.79%) 1	0.459
Race White Black/African American Asian American Indian/Alaskan Native Native Hawaiian/Pacific Islander Other Missing	 11702 (94.16%) 487 (3.92%) 37 (0.30%) 30 (0.24%) 9 (0.07%) 163 (1.31%) 19	 226 (93.39%) 9 (3.72%) 0 (0.00%) 0 (0.00%) 1 (0.41%) 6 (2.48%)	 11476 (94.17%) 478 (3.92%) 37 (0.30%) 30 (0.25%) 8 (0.07%) 157 (1.29%) 19	0.162
Pre-procedure Creatinine Level (mean, SD) Missing	1.14 ± 0.72 831	1.12 ± 0.44 10	1.15 ± 0.72 821	0.528
Currently On Dialysis Missing	207 (1.67%) 19	4 (1.66%) 1	203 (1.67%) 18	0.994
Tobacco History Current Former Never Missing	 3450 (27.78%) 5558 (44.75%) 3411 (27.47%) 28	 71 (29.46%) 113 (46.89%) 57 (23.65%) 1	 3379 (27.75%) 5445 (44.71%) 3354 (27.54%) 27	0.406
Hypertension Missing	11108 (89.25%) 1	216 (89.26%)	10892 (89.25%) 1	0.997
Dyslipidemia Missing	10127 (81.37%) 1	201 (83.06%)	9926 (81.33%) 1	0.495
Peripheral Arterial Disease Missing	3789 (30.45%) 2	83 (34.30%)	3706 (30.37%) 2	0.188

	Total	Both Vital and NIHSS at F/U		
	n = 12447	1 n = 242	0 n = 12205	P-Value
Diabetes Mellitus Missing	4363 (35.06%) 1	86 (35.54%)	4277 (35.05%) 1	0.873
Ischemic Heart Disease Missing	5259 (42.26%) 3	115 (47.52%)	5144 (42.16%) 3	0.094
History of Heart Failure Missing	1174 (9.44%) 4	22 (9.09%)	1152 (9.44%) 4	0.853
Most Recent LVEF% Missing	57.80 ± 11.26 5804	60.19 ± 9.90 108	57.75 ± 11.28 5696	0.013
History of Atrial Fibrillation or Flutter Missing	1387 (11.16%) 15	23 (9.50%)	1364 (11.19%) 15	0.409
Restenosis in Target Vessel After Prior CAS Missing	21 (0.17%) 3	0 (0.00%)	21 (0.17%) 3	0.518
Restenosis in Target Vessel After Prior CEA Missing	219 (1.76%) 2	5 (2.07%)	214 (1.75%) 2	0.714
Target Lesion Symptomatic w/in Past 6 Months Missing	4345 (34.92%) 3	89 (36.78%)	4256 (34.88%) 3	0.539
Visible Thrombus Present Missing	854 (6.86%) 5	32 (13.22%)	822 (6.74%) 5	< 0.001
Neurologic history and Risk Factors Prior to Procedure				
Dementia or Alzheimer s Disease Missing	293 (2.35%) 3	3 (1.24%)	290 (2.38%) 3	0.248
History of Seizure or Known Seizure Disorder Missing	250 (2.01%) 6	5 (2.07%)	245 (2.01%) 6	0.949
Neurologic Event(s) Prior to Procedure Missing	5256 (42.24%) 5	108 (44.63%)	5148 (42.20%) 5	0.448
Prior TIA	3220 (25.87%)	64 (26.45%)	3156 (25.86%)	0.836
Prior Ischemic stroke	1690 (13.58%)	28 (11.57%)	1662 (13.62%)	0.357
Prior Hemorrhage or Hemorrhagic Stroke	77 (0.62%)	1 (0.41%)	76 (0.62%)	0.680
Acute Evolving Stroke Missing	0 (0.00%) 25	0 (0.00%)	0 (0.00%) 25	
Neurologic Status Pre-procedure				
Pre-procedure NIH Stroke Scale Total Score Missing	1.02 ± 2.46 11295	0.51 ± 1.74 86	1.09 ± 2.54 11209	0.005
Pre-procedure Modified Rankin Score Missing	0.31 ± 0.77 10856	0.24 ± 0.60 129	0.32 ± 0.78 10727	0.283
Pre Procedural Meds				
Pre-procedure Aspirin No Yes Contraindicated Missing	 3976 (31.97%) 8329 (66.97%) 131 (1.05%) 11	 60 (24.79%) 180 (74.38%) 2 (0.83%)	 3916 (32.11%) 8149 (66.83%) 129 (1.06%) 11	0.046
Pre-procedure Clopidogrel No Yes Contraindicated Missing	 9699 (78.01%) 2658 (21.38%) 76 (0.61%) 14	 195 (80.58%) 46 (19.01%) 1 (0.41%)	 9504 (77.96%) 2612 (21.43%) 75 (0.62%) 14	0.601

	Total	Both Vital and NIHSS at F/U		
	n = 12447	1 n = 242	0 n = 12205	P-Value
Pre-procedure Ticlopidine No Yes Contraindicated Missing	 12411 (99.82%) 13 (0.10%) 9 (0.07%) 14	 242 (100.00%) 0 (0.00%) 0 (0.00%)	 12169 (99.82%) 13 (0.11%) 9 (0.07%) 14	0.803
Intra Procedure Meds				
Intra-procedure Unfractionated Heparin No Yes Contraindicated Missing	 443 (3.56%) 11984 (96.36%) 10 (0.08%) 10	 9 (3.72%) 233 (96.28%) 0 (0.00%)	 434 (3.56%) 11751 (96.36%) 10 (0.08%) 10	0.897
Intra-procedure LMWH No Yes Contraindicated Missing	 12315 (99.03%) 118 (0.95%) 2 (0.02%) 12	 239 (98.76%) 3 (1.24%) 0 (0.00%)	 12076 (99.04%) 115 (0.94%) 2 (0.02%) 12	0.877
Intra-procedure Thrombin Inhibitors (Any) Missing	64 (0.51%) 12	1 (0.41%)	63 (0.52%) 12	0.823
Post Procedure Meds				
Post-procedure Unfractionated Heparin No Yes Contraindicated Missing	 11522 (92.71%) 889 (7.15%) 17 (0.14%) 19	 212 (87.60%) 30 (12.40%) 0 (0.00%)	 11310 (92.81%) 859 (7.05%) 17 (0.14%) 19	0.005
Post-procedure LMWH No Yes Contraindicated Missing	 11916 (95.89%) 494 (3.98%) 17 (0.14%) 20	 223 (92.15%) 19 (7.85%) 0 (0.00%)	 11693 (95.96%) 475 (3.90%) 17 (0.14%) 20	0.006
Discharge Meds				
Aspirin at Discharge No Yes Contraindicated Missing	 1475 (11.89%) 10838 (87.35%) 94 (0.76%) 40	 26 (10.74%) 213 (88.02%) 3 (1.24%)	 1449 (11.91%) 10625 (87.34%) 91 (0.75%) 40	0.594
Clopidogrel at Discharge No Yes Contraindicated Missing	 7948 (64.15%) 4376 (35.32%) 66 (0.53%) 57	 134 (55.37%) 106 (43.80%) 2 (0.83%)	 7814 (64.32%) 4270 (35.15%) 64 (0.53%) 57	0.015
Ticlopidine at Discharge No Yes Contraindicated Missing	 12343 (99.72%) 21 (0.17%) 14 (0.11%) 69	 240 (99.17%) 0 (0.00%) 2 (0.83%)	 12103 (99.73%) 21 (0.17%) 12 (0.10%) 69	0.003
Warfarin at Discharge No Yes Contraindicated Missing	 11306 (91.31%) 1050 (8.48%) 26 (0.21%) 65	 222 (91.74%) 18 (7.44%) 2 (0.83%)	 11084 (91.30%) 1032 (8.50%) 24 (0.20%) 65	0.091

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

2012 data were used to assess the prevalence of complete follow-up. All available data from 2007-2013 were used to assess the test-retest reliability of the data elements used to describe patient characteristics. We also

restricted the test-retest cohort to those hospitals that performed >30 procedures over this time period to provide more reliable estimates of the reproducibility of these data elements.

1.8 What were the patient-level sociodemographic (SDS) variables that were available and analyzed in the data or sample used? For example, patient-reported data (e.g., income, education, language), proxy variables when SDS data are not collected from each patient (e.g. census tract), or patient community characteristics (e.g. percent vacant housing, crime rate).

The race of the patients was used in the analysis of the data.

2a2. RELIABILITY TESTING

***Note**: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

Level	Signal-to-Noise
All, >10 Procedures	.982
>Q1 (>77 Procedures)	.985
>Q2 (>166 Procedures)	.994
>Q3 (>315 Procedures)	.995
>Average (>234 Procedures)	.996

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Among centers reporting greater than 30 cases (n=XX) between years 2007-2013, the accuracy of data elements entered into the CARE registry was compared. This approach enabled us to examine 2 independent abstractions of data for the same patient. For characteristics that generally do not change (e.g. gender), we would expect near perfect reproducibility. For other characteristics (e.g. diabetes) we would expect that any patient diagnosed with diabetes on the first visit should also have diabetes recorded on the second visit. It is, however, plausible that someone could be diagnosed with diabetes between their first and second visit, so the emergence of diabetes on the second visit is not necessarily an 'error' and no interpretation is made for these scenarios.

There were 449 patients in the CARE registry that had 2 procedures between 2007-2013. Important data elements, support the overall validity of the registry, are provided below:

Age, as defined by date of birth, did not differ in any of the cases.

Gender did not vary in any of the patient records.

Race did not vary in any of the records.

Smoking had minimal inconsistencies. There were 3 patients (0.67%) who were categorized as current smokers on their 1st procedure and never smokers on their 2nd procedure. There were 9 patients (2.0%) listed as former smokers on their 1st visit and never smokers on their 2nd procedure.

History of Peripheral Artery Disease was noted in 11 patients (2.4%) at the time of their 1st procedure, but not at the time of their 2nd.

Diabetes was noted in 6 patients (1.3%) at the time of their 1st procedure, but not at the time of their 2nd.

Chronic Lung Disease was noted in 7 patients (1.6%) at the time of their 1st procedure, but not at the time of their 2nd.

Dyslipidemia was noted in 12 patients (2.7%) at the time of their 1st procedure, but not at the time of their 2nd.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)
For the elements that we were able to assess, we believe that the results reported in the NCDR CARE record are consistent and accurate.

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☐ **Performance measure score**

- ☐ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*
Face Validity of this outcome- As expressed in the application, this measure was shared with a noted panel of expert cardiologist who participate regularly in carotid artery stenting procedures. All members of the committee reported that the measure appears to be a good indication of a positive outcome in carotid artery stenting procedures.

Content validity of this outcome – post-procedure stroke is one of the major adverse outcomes from carotid artery stenting and carotid endarterectomy. For example, this was the major outcome in the recent CREST trial, a randomized comparison of carotid artery stenting and surgical endarterectomy. A recent meta-analysis by Bangalore and colleagues (Arch Nerol 201; 68:172-84) summarized the results of 13 randomized controlled trials to assess the comparative effectiveness of stenting vs. surgery for carotid occlusion on the outcomes of

myocardial infarction, death and stroke. The latter 2 of these outcomes are proposed to serve as a process measure for quantifying the quality of carotid revascularization by this measure.

There is a sound clinical rationale for systematically measuring the outcomes of carotid revascularization. First, without knowing the outcomes, a hospital cannot know if it is applying its treatment in a safe and effective manner. Given how infrequently current providers assess the 30-day survival and stroke outcomes, it is obvious that more than half of these hospitals have no foundation with which to assess the quality of their care. We have proposed a process measure, merely assessing the stroke-free survival of treated patients, because without clearer ascertainment of outcomes it is not possible to provide risk-adjusted comparisons across centers and provide clear benchmarks of performance to identify hospitals that have the opportunity to improve. Second, as the country seeks to support the use of evidence-based medicine, the majority of the evidence in carotid disease comes from clinical trials. However, many of the trials establishing the benefits of carotid revascularization require that centers document a certain success rate, without complications of stroke or death, before the center can participate in a clinical trial. If a center does not know its rate, it will not know whether or not the benefits observed in a clinical trial apply to their practice. Finally, for patients to be adequately informed about the risks and benefits of treatment, hospitals need to have reliable data to share with their patients. By collecting, analyzing and reporting the outcomes of treatment, hospitals will be much better able to provide their patients the information that they need to make a treatment decision.

In developing this measure, the ACC consulted with leading experts in the field and vetted the process measure with the following committees. The individuals within specific committees and workgroups are noted below:

NCDR Strategic Quality and Oversight Committee— an ACC leadership oversight committee that serves as the primary resource for crosscutting scientific and quality of care methodological issues – ensured the data dictionaries and metrics are consistent across registries. They also reviewed and approved the methodology and results of the bleeding outcome and model.
These members include Dr. Frederick Masoudi (chair), Dr. David Malenka, Dr. Thomas Tsai, Dr. Matthew Reynolds,Dr. David Shahian, Dr. John Windle, Dr. Fred Resnic, Dr. John Moore, Dr. Deepak Bhatt, Dr. James Tcheng, Dr. Jeptha Curtis, Dr. Paul Chan, Dr. Matt Roe, and Dr. John Rumsfeld

NCDR Clinical Workgroup is a designated set of experts that oversees this NQF application. Prior to submission, it ensures there is variation in care, disparities data, and that the measure is a true reflection of quality care at a particular site and can also be used to improve quality. This committee included Dr. Jeptha Curtis (chair), Dr. Frederick Masoudi, Dr. John Rumsfeld, Dr. Christopher White, and Dr. Thomas Tsai.

NCDR CARE/PVI Transition Committee provides strategic direction for the Registry and ensures the measures submitted to NQF met key criterion such as reliability, feasibility, and that there is compelling evidence base behind the development and implementation of this measure, which included Christopher White (Chair), Kalon Ho, Ken Rosenfield, Bobby Yeh, Michael Jaff, Thomas Tsai, P. Michael Grossman, Herbert Aronow, H. Vernon Anderson

Lastly the 16 member NCDR Management Board and 31member ACCF Board of Trustees approved these measures for submission to NQF.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)
As this measure is being proposed primarily on the basis of its content validity, as described above, there are no empiric results from formal validity testing.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

As described above, we believe that acquiring the short-term outcomes of carotid revascularization is a critical foundation for assessing and improving care. The CARE registry provides the infrastructure to enter and analyze these data, if collected. An approved performance measure will increase the acquisition of these data and enable quality to be assessed and improved.

2b3. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

The only proposed exclusion is for patients being treated in the context of an acute evolving stroke. This is a distinct clinical setting from the treatment of stable carotid disease. Moreover, the neurological results are likely to be strongly influenced by the presenting stroke, more so than the revascularization procedure that they receive.

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

The Table below shows the differences in clinical characteristics of those with an acute evolving stroke as compared with those treated for stable carotid disease:

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

Given the 2 very distinct populations, we believe that a measure of the survival and neurological outcomes at 30 days is an internally consistent, clinically-interpretable measure that does not suffer from excluding those with an acute ischemic stroke.

ex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	12447	96.89	12447	96.89
AES	123	0.96	12570	97.84
Spont Dis	28	0.22	12598	98.06
Hospitals performing >30 procedures	249	1.94	12847	100.00

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?
☒ **No risk adjustment or stratification**
☐ **Statistical risk model with** Click here to enter number of factors **risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. 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 (case mix) is not needed to achieve fair comparisons across measured entities.

Not applicable.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or sociodemographic factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*)

Not applicable.

2b4.4a. What were the statistical results of the analyses used to select risk factors?

2b4.4b. Describe the analyses and interpretation resulting in the decision to select SDS factors (e.g. prevalence of the factor across measured entities, empirical association with the outcome, contribution of unique variation in the outcome, assessment of between-unit effects and within-unit effects)
Not applicable.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)
Not applicable.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**
Not applicable.

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):
Not applicable.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Not applicable.

2b4.9. Results of Risk Stratification Analysis:
Not applicable.

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for*

the test conducted)
Not applicable.

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)
Not applicable.

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

We observed marked variation in the collection of 30-day outcomes data among hospitals performing at least 30 procedures in 2012. Among 180 hospitals performing carotid revascularization in 17,289 patients in 2012, the range of hospital's collection of 30-day outcomes varied from 0% to 50%. The interquartile ranges were 0-3.2%, 3.2-26.6%, 26.6-59.3% and 59.3-100%. The variation in patient characteristics, by quartile, is provided below:

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a "zero" f/u rate n = 8304	Patients in hospitals with "any" follow-up rate. n = 4143	P-Value
Followup Measures				
Alive or deceased recorded	7241 (58.17%)	4370 (52.63%)	2871 (69.30 %)	< 0.001
NIHSS recorded	242 (1.94%)	0 (0.00%)	242 (5.84%)	< 0.001
A. Demographics				
Age	70.78 ± 10.58	70.80 ± 10.64	70.75 ± 10.47	0.832
Sex Male Female Missing	 7375 (59.26%) 5071 (40.74%) 1	 4927 (59.34%) 3376 (40.66%) 1	 2448 (59.09 %) 1695 (40.91 %)	0.787

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a “zero” f/u rate n = 8304	Patients in hospitals with “any” follow-up rate. n = 4143	P-Value
Race				< 0.001
White	11702 (94.16%)	7805 (94.18%)	3897 (94.11 %)	
Black/African American	487 (3.92%)	356 (4.30%)	131 (3.16%)	
Asian	37 (0.30%)	24 (0.29%)	13 (0.31%)	
American Indian/Alaskan Native	30 (0.24%)	17 (0.21%)	13 (0.31%)	
Native Hawaiian/Pacific Islander	9 (0.07%)	4 (0.05%)	5 (0.12%)	
Other	163 (1.31%)	81 (0.98%)	82 (1.98%)	
Missing	19	17	2	
Pre-procedure Creatinine Level	1.14 ± 0.72	1.15 ± 0.74	1.14 ± 0.66	0.510
Missing	831	695	136	
Currently On Dialysis	207 (1.67%)	146 (1.76%)	61 (1.47%)	0.239
Missing	19	13	6	
Tobacco History				0.090
Current	3450 (27.78%)	2267 (27.38%)	1183 (28.57 %)	
Former	5558 (44.75%)	3689 (44.56%)	1869 (45.14 %)	
Never	3411 (27.47%)	2323 (28.06%)	1088 (26.28 %)	
Missing	28	25	3	
Hypertension	11108 (89.25%)	7444 (89.65%)	3664 (88.44 %)	0.039
Missing	1	1		
Dyslipidemia	10127 (81.37%)	6693 (80.61%)	3434 (82.89 %)	0.002
Missing	1	1		
Peripheral Arterial Disease	3789 (30.45%)	2532 (30.49%)	1257 (30.35 %)	0.876
Missing	2		2	
Diabetes Mellitus	4363 (35.06%)	2913 (35.08%)	1450 (35.01 %)	0.936
Missing	1		1	
Ischemic Heart Disease	5259 (42.26%)	3517 (42.36%)	1742 (42.06 %)	0.744
Missing	3	2	1	
History of Heart Failure	1174 (9.44%)	742 (8.94%)	432 (10.43%)	0.007
Missing	4	4		

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a “zero” f/u rate n = 8304	Patients in hospitals with “any” follow-up rate. n = 4143	P-Value
Most Recent LVEF% Missing	57.80 ± 11.26 5804	57.94 ± 11.26 3855	57.52 ± 11.27 1949	0.161
History of Atrial Fibrillation or Flutter Missing	1387 (11.16%) 15	932 (11.24%) 13	455 (10.99%) 2	0.672
Restenosis in Target Vessel After Prior CAS Missing	21 (0.17%) 3	13 (0.16%) 2	8 (0.19%) 1	0.639
Restenosis in Target Vessel After Prior CEA Missing	219 (1.76%) 2	138 (1.66%) 2	81 (1.96%)	0.241
Target Lesion Symptomatic w/in Past 6 Months Missing	4345 (34.92%) 3	2661 (32.06%) 3	1684 (40.65%)	< 0.001
Visible Thrombus Present Missing	854 (6.86%) 5	531 (6.40%) 5	323 (7.80%)	0.003
Neurologic History and Risk Factors Pre-procedure				
Dementia or Alzheimer s Disease Missing	293 (2.35%) 3	201 (2.42%) 3	92 (2.22%)	0.486
History of Seizure or Known Seizure Disorder Missing	250 (2.01%) 6	144 (1.74%) 6	106 (2.56%)	0.002
Neurologic Event(s) Prior to Procedure Missing	5256 (42.24%) 5	3304 (39.81%) 5	1952 (47.12%)	< 0.001
Prior TIA	3220 (25.87%)	1989 (23.95%)	1231 (29.71%)	< 0.001
Prior Ischemic stroke	1690 (13.58%)	1090 (13.13%)	600 (14.48%)	0.037
Prior Hemorrhage or Hemorrhagic Stroke	77 (0.62%)	49 (0.59%)	28 (0.68%)	0.565
Acute Evolving Stroke Missing	0 (0.00%) 25	0 (0.00%) 14	0 (0.00%) 11	
Neurologic Status Pre-procedure				

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a "zero" f/u rate n = 8304	Patients in hospitals with "any" follow-up rate. n = 4143	P-Value
Pre-procedure NIH Stroke Scale Total Score Missing	1.02 ± 2.46 11295	2.39 ± 4.65 8185	0.86 ± 2.01 3110	< 0.001
Pre-procedure Modified Rankin Score Missing	0.31 ± 0.77 10856	0.28 ± 0.75 7492	0.35 ± 0.79 3364	0.081
Pre Procedural Meds				
Pre-Procedure Aspirin No Yes Contra Missing	 3976 (31.97%) 8329 (66.97%) 131 (1.05%) 11	 2809 (33.86%) 5397 (65.06%) 90 (1.08%) 8	 1167 (28.19 %) 2932 (70.82 %) 41 (0.99%) 3	< 0.001
Pre-Procedure Clopidogrel No Yes Contra Missing	 9699 (78.01%) 2658 (21.38%) 76 (0.61%) 14	 6458 (77.85%) 1788 (21.56%) 49 (0.59%) 9	 3241 (78.32 %) 870 (21.02%) 27 (0.65%) 5	0.735
Pre-Procedure Ticlopidine No Yes Contra Missing	 12411 (99.82%) 13 (0.10%) 9 (0.07%) 14	 8277 (99.78%) 10 (0.12%) 8 (0.10%) 9	 4134 (99.90 %) 3 (0.07%) 1 (0.02%) 5	0.271
Intra Procedure Meds				
Intra-Procedure Unfractionated Heparin No Yes Contra Missing	 443 (3.56%) 11984 (96.36%) 10 (0.08%) 10	 328 (3.95%) 7961 (95.94%) 9 (0.11%) 6	 115 (2.78%) 4023 (97.20 %) 1 (0.02%) 4	0.001

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a "zero" f/u rate n = 8304	Patients in hospitals with "any" follow-up rate. n = 4143	P-Value
Intra-Procedure LMWH No Yes Contra Missing	12315 (99.03%) 118 (0.95%) 2 (0.02%) 12	8188 (98.70%) 107 (1.29%) 1 (0.01%) 8	4127 (99.71%) 11 (0.27%) 1 (0.02%) 4	< 0.001
Intra-Procedure Thrombin inhibitors (Any) Missing	64 (0.51%) 12	51 (0.61%) 8	13 (0.31%) 4	0.027
Post Procedure Meds				
Post-Procedure Unfractionated Heparin No Yes Contra Missing	11522 (92.71%) 889 (7.15%) 17 (0.14%) 19	7642 (92.19%) 642 (7.75%) 5 (0.06%) 15	3880 (93.74%) 247 (5.97%) 12 (0.29%) 4	< 0.001
Post-Procedure LMWH No Yes Contra Missing	11916 (95.89%) 494 (3.98%) 17 (0.14%) 20	7930 (95.69%) 357 (4.31%) 0 (0.00%) 17	3986 (96.28%) 137 (3.31%) 17 (0.41%) 3	< 0.001
Discharge Meds				
Aspirin at Discharge No Yes Contra Missing	1475 (11.89%) 10838 (87.35%) 94 (0.76%) 40	1050 (12.67%) 7168 (86.52%) 67 (0.81%) 19	425 (10.31%) 3670 (89.03%) 27 (0.66%) 21	< 0.001
Clopidogrel at Discharge No Yes Contra Missing	7948 (64.15%) 4376 (35.32%) 66 (0.53%) 57	5328 (64.42%) 2903 (35.10%) 40 (0.48%) 33	2620 (63.61%) 1473 (35.76%) 26 (0.63%) 24	0.417

	Total	Follow-up Rate		
	n = 12447	Patients in hospitals with a “zero” f/u rate n = 8304	Patients in hospitals with “any” follow-up rate. n = 4143	P-Value
Ticlopidine at Discharge				0.170
No	12343 (99.72%)	8236 (99.66%)	4107 (99.83%)	
Yes	21 (0.17%)	18 (0.22%)	3 (0.07%)	
Contra	14 (0.11%)	10 (0.12%)	4 (0.10%)	
Missing	69	40	29	
Warfarin at Discharge				0.596
No	11306 (91.31%)	7537 (91.16%)	3769 (91.61%)	
Yes	1050 (8.48%)	712 (8.61%)	338 (8.22%)	
Contra	26 (0.21%)	19 (0.23%)	7 (0.17%)	
Missing	65	36	29	

The variations in hospital characteristics, stratified by quartile of follow-up data collection, is shown below:

		fu_rate		
	n = 53	0.0000000000 to <.000001 n = 38	.000001 to 0.4453780000 n = 15	P-Value
Teaching Hospital	19 (35.85%)	15 (39.47%)	4 (26.67%)	0.381
Public Hospital	29 (54.72%)	20 (52.63%)	9 (60.00%)	0.627
Type of Hospital				0.525
Government	52 (98.11%)	37 (97.37%)	15 (100.00%)	
Private/Community	1 (1.89%)	1 (2.63%)	0 (0.00%)	
University				

		fu_rate		
	n = 53	0.0000000000 to <.000001 n = 38	.000001 to 0.4453780000 n = 15	P-Value
Location of Hospital Rural Suburban Urban	8 (15.09%) 24 (45.28%) 21 (39.62%)	7 (18.42%) 18 (47.37%) 13 (34.21%)	1 (6.67%) 6 (40.00%) 8 (53.33%)	0.348
Procedural Volume in 2012	234.85 ± 212.05	218.53 ± 216.82	276.20 ± 200.56	0.377

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

We observed that patient characteristics of hospitals in the lower performance quartiles or outcomes reporting did not differ substantially from a clinical perspective, other than that their patients were more likely to be symptomatic within the past 6 months with worse NIHSS and modified Rankin Scores pre-procedurally.

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)
Hospital level performance variation with respect to this metric varied from 0% to 100%The acquisition of outcomes data by hospitals after performing carotid revascularization could not be more broad ranging, from never to sometines assessing patients' outcomes. As described in Section 2b2.2, knowing and understanding a hospital's performance is essential for providing safe, evidence-based, patient-centered care. Importantly, since some hospitals were able to assess the survival and neurological outcomes of all of their patients, it is currently feasible to do so.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

If only one set of specifications, this section can be skipped.

Note: *This item is directed to measures that are risk-adjusted (with or without SDS factors)* ***OR*** *to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***Comparability is not required when comparing performance scores with and without SDS 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.***

2b6.1. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

We are not proposing alternative methods for data collection or performance assessment.

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

2b6.3. What is your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

No empirical analysis was performed. However, it was felt that the method employed would minimize the potential for gaming.

The measure is specified such that cases with missing data are assumed to have not met the metric. The performance ranges throughout this application reflect this approach. By following this method, the scores should be a true depiction of performance scores.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

See tables above for frequency of missing data.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

Given the low frequency of exclusions, we do not believe that the exclusions have any impact on the validity, accuracy or interpretability of this measure. The exclusions have little potential for bias especially given the CARE Data Quality Program audits all essentially performance measure elements on a 3 year cycle and would detect misclassifications of patient records.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

ALL data elements are in defined fields in electronic clinical data (e.g., clinical registry, nursing home MDS, home health OASIS)

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

Participating hospitals report patient demographics, medical history, risk factors, hospital presentation, initial cardiac status, procedural details, medications, laboratory values and in-hospital outcomes. The majority of the required data elements are routinely generated and acquired during the delivery of care to this patient population. Electronic extraction of data recorded as part of the procedure expedites data collection. This strategy offers point of care collection and minimizes time and cost. Institutions can manually report using a free web-based tool or automate the reporting by using certified software developed by third-party vendors. The data elements required for this measure are readily available within the patient's medical record or can be attained without undue burden within the hospital. Most data elements exist in a structured format within patient's electronic health record.

The NCDR Data Quality Program consists of 3 main components: data completeness, consistency, and accuracy. Completeness focuses on the proportion of missing data within fields, whereas consistency determines the extent to which logically related fields contain values consistent with other fields. Accuracy characterizes the agreement between registry data and the contents of original charts from the hospitals submitting data.

The Data Quality Report (DQR) consists of registry-specific algorithms that require predetermined levels of completeness and consistency for submitted data fields. Before entering the Enterprise Data Warehouse (EDW), all submissions are scored for file integrity and data completeness, receiving 1 of 3 scores that are transmitted back to facilities using a color coding scheme. A "red

light" means that a submission has failed because of file integrity problems such as excessive missing data and internally inconsistent data. Such data are not processed or loaded into the EDW. A "yellow light" status means that a submission has passed the integrity checks but failed in completeness according to predetermined thresholds. Such data are processed and loaded into the EDW but are not included in any registry aggregate computations until corrected. Facilities are notified about data submission problems and provided an opportunity to resubmit data. Finally, a "green light" means that a submission has passed all integrity and quality checks. Such submissions are loaded to the EDW. After passing the DQR, data are loaded into a common EDW that houses data from all registries and included for all registry aggregate computations. In a secondary transaction process, data are loaded into registry-specific, dimensionally modeled data marts.

The conditions of participation in the NCDR CARE Registry requires inclusion of consecutive patients; thus the registry is designed to include all adult patients who undergo a carotid revascularization procedure at participating institutions. Section 2.b of the NCDR Master Agreement with participants includes 'Participant Responsibilities': "b. Use of ACCF Data Set and ACCF-Approved Software. Participant will submit a data record on each patient who receives medical care and who is eligible for inclusion in the Registries in which Participant is participating under this Agreement." Adult patients, ages 18 years and older, who have a carotid revascularization procedure. Patients are selected for inclusion by reviewing existing medical records and no direct interaction with the patient will be required outside of the normal course of care. There will be no discrimination or bias with respect to inclusion on the basis of sex, race, or religion.
Patient confidentiality is preserved in reports as the data are aggregated for the purposes of assessing performance. The CARE dataset, comprised of approximately 250, data elements was created by a panel of experts using available ACC-AHA guidelines, data elements and definitions, and other evidentiary sources. Private health information (PHI), such as social security number, is collected. The intent for collection of PHI is to allow for registry interoperability and the potential for future generation of patient-level drill downs in Quality and Outcomes Reports. Registry sites can opt out of transmitting direct identifiers to the NCDR, however, so inclusion of direct identifiers in the registry is at the discretion of the registry participants themselves. When using the NCDR web-based data collection tool, direct identifiers are entered but a partition between the data collection process and the data warehouse maintains the direct identifiers separate from the analysis datasets. The minimum level of PHI transmitted to the ACCF when a participant opts out of submitting direct identifiers meets the definition of a Limited Dataset as such term is defined by the Health Insurance Portability and Accountability Act of 1996.
Data collection within the NCDR conforms to laws regarding protected health information. Patient confidentiality is of utmost concern with all metrics. The proposed measure does not include a patient survey. Physician and/or institutional confidentiality CARE Registry. No testing, time, risk, or procedures beyond those required for routine care will be imposed. The primary risk associated with this measure is the potential for a breach of patient confidentiality. The ACCF has established a robust plan for ensuring appropriate and commercially reasonable physical, technical, and administrative safeguards are in place to mitigate such risks.

Data are maintained on secure servers with appropriate safeguards in place. The project team periodically reviews all activities involving protected health information to ensure that such safeguards including standard operating procedures are being followed. The procedure for notifying the ACCF of any breach of confidentiality and immediate mitigation standards that need to be followed is communicated to participants. ACCF limits access to Protected Health Information, and to equipment, systems, and networks that contain, transmit process or store Protected Health Information, to employees who need to access the PHI for purposes of performing ACCF's obligations to participants who are in a contractual relationship with the ACCF. All PHI are stored in a secure facility or secure area within ACCF's facilities which has separate physical controls to limit access, such as locks or physical tokens. The secured areas are monitored 24 hours per day, 7 days per week, either by employees or agents of ACCF by video surveillance, or by intrusion detection systems.

Each participant who has access to the NCDR website must have a unique identifier. The password protected webpages have implemented inactivity time-outs. Encryption of wireless network data transmission and authentication of wireless devices containing NCDR Participant's information ACCF's network is required. Protected Health Information may only be transmitted off of ACCF's premises to approved parties, which shall mean: A subcontractor who has agreed to be bound by the terms of the Business Associate Agreement between the ACCF and the NCDR Participant.

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.
The ACCF's program the National Cardiovascular Data Registry (NCDR) provides evidence based solutions for cardiologists and other medical professionals committed to excellence in cardiovascular care. NCDR hospital participants receive confidential benchmark reports that include access to measure macro specifications and micro specifications, the eligible patient population, exclusions, and model variables (when applicable). In addition to hospital sites, NCDR Analytic and Reporting Services provides consenting hospitals' aggregated data reports to interested federal and state regulatory agencies, multi-system provider groups, third-party payers, and other organizations that have an identified quality improvement initiative that supports NCDR-participating facilities. Lastly, the ACCF

also allows for licensing of the measure specifications outside of the Registry. For calendar year 2016 the annual pricing for hospitals, NCDR Analytic and Reporting Services, and licensing of measure specifications is $4,530.00.
Measures that are aggregated by ACCF and submitted to NQF are intended for public reporting and therefore there is no charge for a standard export package. However, on a case by case basis, requests for modifications to the standard export package will be available for a separate charge.
There is no added procedural risk to patients through their hospital's involvement in the CARE Registry. No testing, time, risk, or procedures beyond those required for routine care will be imposed.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

CARE Registry of the National Cardiovascular Data Registry of the American College of Cardiology

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
Plan is to publically report in the future.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)
ACC is committed to implementing this measure. ACC is an authorized organization to receive CMS data through the ResDAC application process. Unfortunately, it has been determined by ResDAC that this authorization does not permit use of CMS for performance measure reporting purposes, either to hospitals or for public display. ACC is currently in process of applying to be a Qualified Entity. It is unclear if this pathway will permit measure implementation. ACC also is commenting on and tracking proposed language in 21st Century Cures legislation, which does appear to create a pathway for use of CMS data for this type of reporting purpose.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Not available, initial endorsement

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

There were no unintended consequences identified.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

2396 : Carotid artery stenting: Evaluation of Vital Status and NIH Stroke Scale at Follow Up

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

N/A

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**

Are the measure specifications completely harmonized?

Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

The measure is specified identically with Measure 2396 Carotid artery stenting: Evaluation of Vital Status and NIH Stroke Scale at Follow Up with the exception of the CAS versus the CEA population. The CAS and CEA populations were previously submitted to NQF as one measure by the American College of Cardiology. However NQF specifically asked that the two measures be uncoupled and submitted separately.

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);

OR

Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
No competing measures.

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** CARE_v109_CEA_DataDictionary.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American College of Cardiology
Co.2 **Point of Contact:** Penelope, Solis, comment@acc.org, 202-375-6576-
Co.3 **Measure Developer if different from Measure Steward:** American College of Cardiology
Co.4 **Point of Contact:** Traci, Connolly, tconnoll@acc.org, 202-375-6298-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
SQOC—Leadership committee that oversaw broad issues and approved submission of given metric to NQF.
Fred Masoudi, David Malenka, Thomas Tsai, Matt Reynolds, David Shahian, John Windle, Fred Resnic, John Moore, Deepak Bhatt, James Tcheng, Jeptha Curtis, Paul Chan, Matt Roe, John Rumsfeld
Clinical SubWorkgroup-oversaw NQF application components
Jeptha Curtis-chair
Christopher White, Thomas Tsai, John Rumsfeld, Fred Masoudi
CARE/PVI Transition Workgroup -Provides strategic direction for the Registry and monitors research and clinical activities.
Chris White, Kalon Ho, Ken Rosenfield, Bobby Yeh, Michael Jaff, Thomas Tsai, P. Michael Grossman, Herb Aronow, H. Vernon Anderson

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:**
Ad.3 **Month and Year of most recent revision:**
Ad.4 **What is your frequency for review/update of this measure?**
Ad.5 **When is the next scheduled review/update for this measure?**

Ad.6 **Copyright statement:** American College of Cardiology Foundation All Rights Reserved
Ad.7 **Disclaimers:** ACC realizes the various NCDR endorsed measures are not readily available on their own main webpage. However, ACCF plans to update their main webpage (acc.org) to include the macrospecifications of the NQF endorsed measures. ACC hopes to work collaboratively with NQF to create a consistent and standard format would be helpful for various end users. In the interim, the supplemental materials include the details needed to understand this model. In addition, interested parties are always able to contact comment@acc.org to reach individuals at the ACC Quality Measurement Team.

Ad.8 **Additional Information/Comments:**

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3030
Measure Title: STS Individual Surgeon Composite Measure for Adult Cardiac Surgery
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: The STS Individual Surgeon Composite Measure for Adult Cardiac Surgery includes five major procedures (isolated CABG, isolated AVR, AVR+CABG, MVRR, MVRR+CABG) and comprises the following two domains:
Domain 1 – Risk-Adjusted Operative Mortality
Operative mortality is defined as death before hospital discharge or within 30 days of the operation.
Domain 2 – Risk-Adjusted Major Morbidity
Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

All measures are based on audited clinical data collected in the STS Adult Cardiac Surgery Database. Individual surgeons with at least 100 eligible cases during the 3-year measurement window will receive a score for each domain and an overall composite score. In addition to calculating composite score point estimates with credible intervals, surgeons will be assigned rating categories designated by the following:
1 star – lower-than-expected performance
2 stars – as-expected performance
3 stars – higher-than-expected performance

Developer Rationale: Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, it fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events.

In addition, with the development of this composite measure, STS addresses a number of major concerns that have previously been raised regarding surgeon-level metrics. It combines results from five of the most frequently performed cardiac surgical procedures, encompassing most of a typical adult cardiac surgeon's practice, as opposed to basing performance on just one or a few separate procedures. Furthermore, it provides a more comprehensive quality

assessment and additional endpoints, as it includes risk-adjusted mortality and the risk-adjusted occurrence of any of five major complications. This measure will be useful to surgeons in identifying potential areas for improvement, and it has numerous advantages compared with existing surgeon metrics if used for accountability purposes.

Numerator Statement: See Appendix
Denominator Statement: See Appendix
Denominator Exclusions: See Appendix

Measure Type: Composite
Data Source: Electronic Clinical Data : Registry
Level of Analysis: Clinician : Individual

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

This measure of healthcare outcomes is comprised of a mortality domain and a morbidity domain, defined as any one or more of the identified complications.

It is based on 12 NQF-endorsed measures of which 7 are mortality measures that the developer identifies are the most frequently performed cardiac surgery procedures encompassing most of a typical adult cardiac surgeon's practice and 5 are cardiac surgery-related major morbidities. The 12 measures are:

- 0119: Risk-Adjusted Operative Mortality for CABG
- 0120: Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)
- 0121: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement
- 0122: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
- 0123: Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery
- 1501: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
- 1502: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery
- 0114: Risk-Adjusted Postoperative Renal Failure
- 0115: Risk-Adjusted Surgical Re-exploration
- 0129: Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
- 0130: Risk-Adjusted Deep Sternal Wound Infection
- 0131: Risk-Adjusted Stroke/Cerebrovascular Accident

- NQF criteria indicate that each component in a composite must meet the evidence subcriterion to justify its inclusion in the composite and be NQF-endorsed or evaluated as meeting measure evaluation criteria.
- The components of this composite are outcomes for which the required evidence is identification of a relationship between the outcome and at least one healthcare action that could achieve change in measure results. Information regarding service and/or care to impact mortality and 4 of the 5 morbidities is provided.
- The NQF-endorsed measures upon which this composite is based are specified for analysis at the group/facility

level.

- References that address operative mortality and morbidity dating from the 1990's through 2014, including those related to current STS adult cardiac surgery risk models, are provided.

Question for the Committee:

- *Is the information regarding development and application of the composite to individual surgeon performance clear and compelling?*
- *Is there at least one thing that providers can do to achieve a change in measure results?*
- *Does the Committee agree that the components together convey an appropriate measure of overall surgeon-specific perioperative care?*

Guidance from the Evidence Algorithm: Assess performance on outcome (Box 1) – Relationship between outcome and healthcare action (Box 2)

Preliminary rating for evidence: X Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- For this new measure, calculation was done using STS data for patients undergoing cardiac surgery during July 2011-June 2014.
  - The developer reports that 9.6% of surgeons with ≥100 cases have lower than expected performance on the measure based on 98% Bayesian credible interval. In comparison, 9.1% of surgeons with ≥10 cases have lower than expected performance.
- Overall results, expressed as estimated distributions of risk-adjusted mortality and morbidity for all procedures and all surgeons in the study cohort – July 1, 2011 to June 30, 2014, were:
  - Mortality – 2.3%; Any major morbidity – 13.7% with prolonged ventilation at 10.2%, deep sternal infection at 0.3%, permanent stroke at 1.4%, renal failure at 2.5% and reoperations at 3.0%.

Disparities

The measure provides information about the performance of surgeons who participate in the STS database. The developer states that there is not a simple way to generate data stratified by patient characteristics at the composite level thus has not presented disparities data. Such data is not required for a new measure.

Questions for the Committee:

- *In considering whether there is a gap in care that warrants a national performance measure, does the fact that each component of the measure represents occurrence of a serious adverse (never) event influence Committee thinking?*
- *Are there areas around which you would like to see disparities information in the future?*

Preliminary rating for opportunity for improvement: ☐ High X Moderate ☐ Low ☐ Insufficient

1c. Composite - Quality Construct and Rationale

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The approach to development of the measure, including decision logic and results of testing (with STS registry

data) used to combine the data into the respective domains (risk-adjusted operative mortality and risk-adjusted major morbidity) and then to combine domain scores into a single composite measure, is presented and described in detail (see Appendix Shahian et al) in a paper that addresses composite measure scoring and provider rating.

- This measure is based on a combination of 7 NQF-endorsed risk-adjusted mortality outcome measures and 5 risk-adjusted major complications (specified for analysis at the group/facility level).
- The composite comprises 2 domains.
- Domain 1 includes risk-adjusted operative mortality (before hospital discharge or within 30 days of operation) for isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG**.**
- Domain 2 includes the occurrence of any one or more of 1) prolonged ventilation; 2) deep sternal wound infection; 3) permanent stroke; 4) renal failure; and 5) reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.
- The developer states that differentiating performance based on mortality alone fails to account for the fact that not all operative survivors received equal quality care. By combining results from 5 of the most frequently performed procedures and risk-adjusted occurrence of any of 5 major complications, it provides a more comprehensive quality assessment that should enable surgeons in identifying potential areas for improvement and has advantages if used for accountability.
- Aggregation and weighting of the composite components is described in the measure information form with greater detail provided in the appendix (see S.4-S.11) and testing form. The mortality domain represents a single outcome though of a number of surgery types; the morbidity domain includes any one of five major morbidities.
- Variables, with definitions (see appendix S.14 and S.15), for CABG, valve, and valve plus CABG risk models are included.
- Mortality and morbidity rates are weighted inversely by respective standard deviations across surgeons.

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*

Preliminary rating for composite quality construct and rationale:

☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

The measure qualifies for a high rating if the Committee determines that the NQF expectation regarding endorsement or evaluation of component measures is satisfied.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- This is a composite measure. The evidence to support development of the composite is clear and compelling. The composite conveys an appropriate measure of overall surgeon care.
- I love this measure
- Again, lower volume surgeons appear to do better. Why?
- Risk adjustment also balances case mix
- 9% lower and 18% higher than expected!

1b.

- Each component of the measure represents a serious adverse event. Generating a composite score will distinguish overall high versus low quality across a range of cardiac surgery services.

1c.

- Yes - clearly states.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The data source for the measure is the STS Adult Cardiac Surgery Database. Data collection occurs through an electronic system using a detailed collection tool and performance reports are provided by STS.
- The measure is specified for analysis at the individual clinician level and intended for use in the hospital/acute care setting.
- The measure is based on a combination of aggregate risk-adjusted operative mortality for isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG and risk-adjusted occurrence of any one or more of 5 major complications to assess STS database participant, individual surgeon performance.
- Surgeon-specific risk-adjusted operative mortality and major complication rates were estimated using a bivariate random-effects logistic regression model. To adjust for case mix, each patient's risk score for operative mortality and his or her risk score for major complications are calculated using existing and modified STS risk models.
- The composite result is calculated (see Shahian et al) as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk adjusted morbidity rate).
- Details regarding development of the risk-adjustment models and the approach to scoring is described in detail in the measure submission and in a 2015 publication.

Questions for the Committee :

- *Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

- The measure was developed and tested using STS data from 2,286 surgeons for patients undergoing cardiac surgery during the 3 year period from July 2011- June 2014.
- A minimum threshold for receiving a surgeon-specific composite score over the three year measurement window was determined after analyzing a number of thresholds in terms of surgeons included, patients included and resulting reliability.
- Estimated reliability of the composite measure using 3 years of data in surgeons with at least 100 cases is reported as 0.81.
- The mathematic approach to signal-to-noise estimation is detailed.

Questions for the Committee:

○ *Is the test sample adequate to generalize for widespread implementation?*

○ *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) – Empiric testing (Box 2) Performance score testing (Box 4) – Method described and appropriate (Box 5) Confidence that scores are reliable (Box 6)

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

○ *Does the Committee agree that the specifications are consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Summarize the validity testing from the prior review: N/A

SUMMARY OF TESTING Method and Results

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- The measure combines surgeon-level operative mortality (death before hospital discharge or within 30 days of operation) for five identified major cardiac procedures and any one or more of 5 specified postoperative major morbidities to reflect individual surgeon-level performance.
  - One time period (July 2011 – June 2014) is used for testing.
  - Degree of uncertainty around a surgeon's composite measure estimate is indicated by calculating 98% Bayesian credible intervals (CI).
  - Point estimates and CI's from an individual surgeon are reported with a comparison to benchmarks (overall average STS composite score and several percentiles) based on the national sample.
  - Also, the composite measure result is converted into one of 3 groups or categories using a Bayesian CI that overlaps the overall STS average as "expected performance". The remaining categories are "lower-than-expected performance" and "higher-than-expected performance".

Validity testing results:

- Among surgeons (1,976) with at least 100 cases from July 2011 – June 2014 (see appendix 1b.2), overall average STS composite scores (with risk-adjusted mortality and risk-adjusted morbidity) were:
  - 1,413 (71.5%) performed as expected (risk adjusted mortality, 2.5% and risk-adjusted morbidity, 14.2%);

- 189 (9.6%) had lower-than-expected performance (risk adjusted mortality, 4.2% and risk-adjusted morbidity, 22.6%); and
- 374 (18.9%) had higher-than-expected performance (risk adjusted mortality 1.2% and risk-adjusted morbidity, 8.8%)

- The developer states that the test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance and differences in morbidity and mortality rates correspond appropriately with changes in performance categories.

Questions for the Committee:

- *Does the Committee agree that ranking scores into 3 performance groups and comparing averages provides validation of the measure such that you can agree that the score from this measure as specified is an indicator of quality?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

No exclusions

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

- The developer states that the performance measure gauges performance of STS surgeons and is not a patient or operation level measure and that it has no simple way to generate data stratified by patient characteristics at the composite level for this surgeon-level measure.
- The developer indicates the method of controlling for difference in case mix includes both a statistical risk model and endorsed (or submitted) individual performance measures.
- A bivariate random-effects logistic regression model was used to estimate surgeon specific risk-adjusted operative mortality and major complication rates.
- Sensitivity analyses were performed with each surgeon's risk-adjusted mortality and complication rates estimated in models that adjust for 41 and 47 individual patient covariates respectively.
- After sensitivity analysis (demonstrating validity of approach), operative mortality risk score and major complication risk score were used as covariates in the hierarchical model for operative mortality and major complication, respectively.
- Risk scores for patients were calculated according to existing published risk models (see articles published 2009). Published models for mitral valve repair and replacement were modified to account for inclusion of patients undergoing tricuspid repair. Details of the estimated odds ratios from the models are included.
  - For the modified MVRR model, the bootstrap-adjusted estimated C-statistic was 0.746 for the morbidity model and 0.807 for the mortality model (comparable to the STS 2008 models when using the same sample - 0.745 and 0807).
  - For the modified MVRR + CABG, the bootstrap-adjusted C-statistic was 0.708 for the morbidity model and 0.738 for the mortality model (comparable to the STS 2008 models when using the same sample – 0.707 and 0.738).
- Discrimination and calibration for the modified models are discussed in terms of the effort to ensure high calibration. Coefficients of each model were re-estimated using the current 3-year study sample and current end point definitions. Risk scores for patients undergoing mitral operation without CABG were calculated using a modified version of the published STS 2008 mortality and major complications models for isolated valve procedures

(*O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.*) to allow inclusion of patients undergoing tricuspid repair.

Questions for the Committee:

- *Are there appropriate risk-adjustment strategies included in the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Do you agree with the developer's rationale that there is no conceptual basis for adjusting this measure for SDS factors?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- As noted in the validity testing section, among surgeons with at least 100 cases over the period July 2011 – June 2014, overall average STS composite scores (with risk-adjusted mortality and risk-adjusted morbidity) were:
  - 1,413 (71.5%) performed as expected (risk adjusted mortality, 2.5% and risk-adjusted morbidity, 14.2%);
  - 189 (9.6%) had lower-than-expected performance (risk adjusted mortality, 4.2% and risk-adjusted morbidity, 22.6%); and
  - 374 (18.9%) had higher-than-expected performance (risk adjusted mortality 1.2% and risk-adjusted morbidity, 8.8%)

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed – single data source.

2b7. Missing Data

- Overall frequency of missing data was 0.4% for operative mortality and 0.3% for major complications.
- The median surgeon-specific frequency of missing data was 0% (range - 0% to 65%) for mortality and 0% (range - 0% to 40%) for the specified major complications.
- The percent of surgeons with >10% missing data was 1.0% for mortality and 1.0% for major complications. *As a sensitivity analysis, the developer recalculated surgeon-specific mortality and complication rates after excluding records with missing data from the denominator. There was high (>0.99) correlation between surgeon-specific rates calculated with missing data excluded versus imputed.*

2d. Composite measure: construction

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

- Pearson correlations were calculated to verify that each of the 2 domains of the measure contribute statistical information but do not dominate the composite. Data from July 2011 – June 2014 were used for the calculation. Results were 0.73 for mortality domain versus overall composite measure and 0.92 for morbidity domain score versus overall score.
- The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus

risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

- Weighting was assessment by an expert panel. It was consistent with the panel's clinical assessment of each domain's relative importance.

Questions for the Committee:

- *Do the component measures fit the quality construct?*
- *Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

Guidance from the Validity Algorithm:
Measure specification consistent with evidence (Box 1) – Potential threats to validity (Box 2) – Empirical validity testing (Box 3) – Face validity testing (Box 4) –
Lack of clarity that analysis provided demonstrates validity = submission insufficient; or
Clarification of validity testing that demonstrates validity = moderate preliminary rating.
Conceptual rationale regarding inclusion of SDS factors needed

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a1.
- All clearly defined.

2b1.
- Totally consistent with evidence

2b2.
- Yes

2b3.
- No

2d.
- Yes

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.
- Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants and based on STS analyses, this includes more than 90% of cardiothoracic programs in the US.
- Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.
- There are no additional costs for data collection specific to the measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

Questions for the Committee:
Is the effort and cost associated with abstracting the required data elements appropriate to the value of the measure?

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 3: Feasibility

All elements are routinely captured in the STS database.

Criterion 4: Usability and Use
Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

This new composite measure was developed in 2014 and published in 2015.

STS plans to make results available to individual surgeons in late 2016 or early 2017. Public reporting will follow within the next several years.

Questions for the Committee*:*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments
Criteria 4: Usability and Use

Planned use in accountability

Criterion 5: Related and Competing Measures

Related or competing measures
Related measures include STS measures that have been included in development of the composite or are otherwise related. They are harmonized.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: Individual Surgeon Composite Measure for Adult Cardiac Surgery
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus
The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related behavior.
- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.
4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or

Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Health outcome: 1. Operative Mortality; 2. Postoperative Major Morbidity

☐ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process:

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Operative Mortality

Mortality likely is the single most important negative outcome associated with a surgical procedure. Operative mortality, defined as death before hospital discharge or within 30 days of the operation, should include nearly all deaths that occur as a direct result of the surgery or an immediate postoperative complication. Critical evaluation of operative mortality allows one to evaluate the risk associated with a given procedure for various patient characteristics, and more importantly, aggressively search for ways to minimize that risk. Preoperative patient selection, surgical timing post coronary event, intraoperative conduct of the case, and many aspects to postoperative care have all been shown to have significant impact on the operative mortality over the last few decades. The published literature (list provided below) on each major procedure included in this composite measure is full of examples of services/care processes that impact operative mortality.

Major Morbidity

- Surgical re-exploration for bleeding remains a known complication following cardiac surgery. The literature documents that bleeding following coronary artery bypass surgery confers greater ICU stay and therefore greater resource consumption. It remains unknown and controversial whether long-term outcomes are worse for the isolated re-exploration for bleeding patients. However, Hein documents that patients with ICU stay > 3 days (with bleeding as multivariate risk factor for this outcome), have a long-term survival which is inferior to patients with ICU stay < 3 days. The patient consequences of this complication relates to the physiological stress of facing another operation and receiving blood products.
- A devastating complication of cardiac surgery is deep sternal wound infection. Patients with deep sternal wound infection require multiple surgeries to clear the infection, have longer hospital stays, greatly increased costs and increased early and late mortality. Care processes that influence the incidence of sternal infection span the first 3 major phases of surgical care. In the preoperative phase, routine patient decontamination and identification of active infections are key elements. In the intraoperative phase, impeccable surgical aseptic technique and timing

of antibiotic administration are key processes. In the postoperative phase, aseptic wound care and detection of early wound inflammation are important in preventing delayed contamination and subsequent infection.

- Prolonged ventilation has been shown to substantially increase length of stay, the costs of care, and is associated with higher rates of respiratory failure, stroke, renal failure, and death. Modalities to decrease the rate of prolonged intubation include physician supervised protocols for extubation implemented by nurses and respiratory therapists, improved preoperative preparation of patients, reduction of postoperative bleeding, and intra-operative protocolized anesthesia care. Current implementation is highly variable and great opportunities to increase the implementation of evidence based care exist. Cardiac surgery programs with high implementation have lower than average rates of prolonged ventilation and significantly lower rates of adverse events.
- Postoperative renal failure is an occasional but serious complication in the cardiac surgical population and is a major determinant of short- and long-term survival. Identification of clinical precursors of postoperative renal insufficiency and improvement in perioperative treatment of this high-risk group will improve the long-term survival of our patients. By implementing known recommendations (delay heart surgery, when possible, after cardiac catheterization, maintain mean CPB perfusion pressure at 80% of preop BP, etc), postoperative kidney injury should be significantly reduced.
- Postoperative stroke/CVA produces significant short- and long-term often devastating effects to patients and their families. It is associated with significant increases in death, respiratory failure, renal failure, length of stay, and cost of care. Modalities to decrease the rate of postoperative stroke include measures to maintain blood pressure and perfusion, glycemic control, avoidance of atrial fibrillation, anticoagulation protocols, etc. Many opportunities exist to decrease stroke rates by increasing implementation of evidence based strategies.

<u>References – Operative Mortality</u>

- Ferguson TB, Hammill BG, et al. A decade of change—risk profiles and outcomes for isolated coronary artery bypass grafting procedures, 1990-1999: a report from the STS National Database Committee and the Duke Clinical Research Institute. *Ann Thorac Surg.* 2002;73(2):480-489; discussion 489-490.
- Grover FL, Shroyer AL, et al. A decade's experience with quality improvement in cardiac surgery using the Veterans Affairs and Society of Thoracic Surgery national databases. *Ann Thorac Surg.*2001; 234(4):464-472; discussion 472-474.
- Hogue CW, Barzilai B, et al. Sex differences in neurologic outcomes and mortality after cardiac surgery: A Society of Thoracic Surgeons National Database report. *Circulation.*2001;03:2133-2137.
- Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg.* 2003;75:1856-1865.
- Williams ML, Muhlbaier LH, Schroder JN, et. al. Risk-adjusted short- and long-term outcomes for on-pump versus off-pump coronary artery bypass surgery. Circulation. 2005 Aug 30;112(9 Suppl):I366-70.
- Shroyer AL, Grover FL, Hattler B, et. al. On-pump versus off-pump coronary artery bypass surgery. N Engl J Med. 2009 Nov 5;361(19):1827-37.
- Hannan EL, Wu C, Smith CR, et. al. Off-pump versus on-pump coronary artery bypass graft surgery: differences in short-term outcomes and in long-term mortality and need for subsequent revascularization. Circulation. 2007 Sep 4;116(10):1145-52. Epub 2007 Aug 20.
- ElBardissi AW, Aranki SF, Sheng S, et al. Trends in isolated coronary artery bypass grafting: an analysis of the Society of Thoracic Surgeons adult cardiac surgery database. J Thorac Cardiovasc Surg. 2012 Feb;143(2):273-81.
- Rangrass G, Ghaferi AA, Dimick JB. Explaining Racial Disparities in Outcomes After Cardiac Surgery: The Role of Hospital Quality. JAMA Surg. 2014;149(3):223-7
- Birkmeyer NJ, Marrin CA, et al. Decreasing mortality for aortic and mitral valve surgery in Northern New England. Northern New England Cardiovascular Disease Study Group. Ann Thorac Surg. 2000;70(2):432-437.

- Edwards FH, Peterson ED, et al. Prediction of operative mortality following valve replacement surgery. JACC. 37:3:885-892.
- Goodney PP, O'Connor GT, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? Ann Thorac Surg. 2003;76:1131-1137.
- Mihaljevic T, Nowicki ER, Rajeswaran J, et. al. Survival after valve replacement for aortic stenosis: implications for decision making. J Thorac Cardiovasc Surg. 2008 Jun;135(6):1270-8; discussion 1278-9.
- Tabata M, Umakanthan R, Cohn LH, et. al. Early and late outcomes of 1000 minimally invasive aortic valve operations. Eur J Cardiothorac Surg. 2008;33(4):537-41.
- Chaliki HP, Mohty D, Avierinos JF, et. al. Outcomes after aortic valve replacement in patients with severe aortic regurgitation and markedly reduced left ventricular function. Circulation. 2002 Nov 19;106(21):2687-93.
- Brennan JM, Holmes DR, Sherwood MW, Edwards FH, et al. The Association of Transcatheter Aortic Valve Replacement Availability and Hospital Aortic Valve Replacement Volume and Mortality in the United States. Ann Thorac Surg. 2014 Dec;98(6):2016-22.
- Thourani VH, Suri RM, et al. Contemporary real-world outcomes of surgical aortic valve replacement in 141,905 low-risk, intermediate-risk, and high-risk patients. Ann Thorac Surg. 2015 Jan;99(1):55-61.
- Chikwe J, Croft LB, Goldstone AB, Castillo JG, Rahmanian PB, Adams DH, et al . Comparison of the results of aortic valve replacement with or without concomitant coronary artery bypass grafting in patients with left ventricular ejection fraction $\leq$30%versus patients with ejection fraction > 30%. Am J Cardiol. 2009;104:1717-21.
- Li Z, Anderson I, Amsterdam EA, Young N, Parker J and Armstrong EJ. Effect of coronary artery disease extent on contemporary outcomes of combined aortic valve replacement and coronary artery bypass graft surgery. Ann Thor Surg 2013;96:2075-82.
- Kobayashi J. Changing strategy for aortic stenosis with coronary artery disease by transcatheter aortic valve implantation. Gen Thorac Cardiovas Surg 2013;61:663-68.
- Beach JM Mihaljevic T, Svensson LG, Rajeswaran J, Marwich T, Griffin B, Johnston DR, Sabik III JF and Blackstone EJ. Coronary artery disease and outcomes of aortic valve replacement for severe aortic stenosis. J Am Coll Cardiol 2013;61:837-48.
- Fukui T, Bando K, Tanaka S, Uchimuro T, Tabata M and Takanashi S. Early and mid-term outcomes of combined aortic valve replacement and coronary artery bypass grafting in elderly patients. Eur J of Cardio-Thorac Surg 2014;45:335-40.
- Mehta RH, Eagle KA, et al. Influence of age on outcomes in patients undergoing mitral valve replacement. Ann Thorac Surg. 2002;74:1459-1467.
- Dayan V, Soca G, et al. Similar survival after mitral valve replacement or repair for ischemic mitral regurgitation: a meta-analysis. Ann Thorac Surg. 2014 Mar;97(3):758-65.
- Kaneko T, Aranki S, et al. Mechanical versus bioprosthetic mitral valve replacement in patients <65 years old. J Thorac Cardiovasc Surg. 2014 Jan;147(1):117-26.
- Iribarne A, Russo MJ, Easterwood R et al. Minimally invasive versus sternotomy approach for mitral valve surgery: a propensity analysis. *Ann Thorac Surg.* 2010;90:1471–1477
- LaPar DJ, Hennessy S, Fonner E, et al. Does urgent or emergent status influence choice in mitral valve operations? An analysis of outcomes from the Virginia Cardiac Surgery Quality Initiative. 2010;90:153-60
- Umakanthan R, Petracek MR, Leacche M et al, Minimally invasive right lateral thoracotomy without aortic cross-clamping: an attractive alternative to repeat sternotomy for reoperative mitral valve surger;y. J Heart Valve Dis. 2010;19:236-43
- Vassileva CM, McNeely C, Spertus J, Markwell S, Hazelrigg S. Hospital volume, mitral repair rates, and mortality in mitral valve surgery in the elderly: An analysis of US hospitals treating Medicare fee-for-service patients. J Thorac Cardiovasc Surg. 2014pii: S0022-5223(14)01290-2
- Chatterjee S, Rankin JS, Gammie JS, et al. Isolated mitral valve surgery risk in 77,836 patients from the Society of Thoracic Surgeons database. Ann Thorac Surg. 2013;96:1587-94

- LaPar DJ, Ailawadi G, Isbell JM, et al. Virginia Cardiac Surgery Quality Initiative. Mitral valve repair rates correlate with surgeon and institutional experience. J Thorac Cardiovasc Surg. 2014;148:995-1003
- Miyata H, Motomura N, Tsukihara H, Takamoto S; Japan Cardiovascular Surgery Database. Risk models including high-risk cardiovascular procedures: clinical predictors of mortality and morbidity. Eur J Cardiothorac Surg. 2010 Nov 1
- Vassileva CM, Boley T, Markwell S, Hazelrigg S. Meta-analysis of short-term and long-term survival following repair versus replacement for ischemic mitral regurgitation. Eur J Cardiothorac Surg. 2010 Aug 18.
- Daneshmand MA, Milano CA, Rankin JS, Honeycutt EF, Shaw LK, Davis RD, Wolfe WG, Glower DD, Smith PK. Influence of patient age on procedural selection in mitral valve surgery. Ann Thorac Surg. 2010 Nov; 90(5):1479-85
- Acker MA, Parides MK, Perrault LP et al (members of Cardiothoracic Surgical Trials Network). Mitral-valve repair versus replacement for severe ischemic mitral regurgitation. N Engl J Med 2014; 370:23-32

References – Major Morbidity

- Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg.* 2003;75:1856-1865.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg.* 2004;77:1137-1139.
- Hein OV, Birnbaum J, Wernecke K, England M, Knoertz W, Spies C. Prolonged Intensive Care Unit Stay in Cardiac Surgery: Risk Factors and Long-Term Survival. *Ann Thor Surg* 2006;81:880-85.
- Karthik S, Grayson AD, McCarron EE, Pullan DM, Desmond MJ. Reexploration for bleeding after coronary artery bypass surgery: risk factors, outcomes, and the effect of time delay. *Ann Thor Surg* 2004;78:527-34.
- Stamou SC, Camp SL, Stiegel RM, et al. Quality improvement program decreases mortality after cardiac surgery. *J Thorac Cardiovasc Surg* 2008;136:494-499.
- Braxton JH, Marrin CA, McGrath PD, et al. 10-Year follow-up of patients with and without mediastinitis. Semin Thorac Cardiovasc Surg. 2004;16:70–76.
- Graf K, Ott E, Vonberg RP, et al. Economic aspects of deep sternal wound infections. Eur J Cardiothorac Surg 2010;37:893-96.
- Speir AM, Kasirajan V, BarnettSD, Fonner E. Additive costs of postoperative complications for isolated coronary artery bypass grafting patients in Virginia. Ann Thorac Surg 2009;88:40-46.
- Olsen MA, Lock-Buckley P, et al. The risk factors for deep and superficial chest surgical site infections after coronary artery bypass graft surgery are different. J Thorac Cardiovasc Surg. 2002;124:136-145.
- Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery; part1 – Conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12.
- Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1 – coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(1 Suppl):S2-S22.
- Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 2 – implementation. Ann Thorac Surg 2011;92:S12-S23.
- Trick WE, Scheckler WE, et al. Modifiable risk factors associated with deep sternal site infection after coronary artery bypass grafting. J Thorac Cardiovasc Surg. 2000;119:108-114.
- Edwards FH, Engelman RM, Houck P et al. The Society of Thoracic Surgeons Practice Guideline Series: Antibiotic Prophylaxis in Cardiac Surgery, Part I: Duration. Ann Thorac Surg 2006; 81: 397 – 404,
- Wilson APL, Gibbons C, Reeves BC, et al. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. BMJ 2004; 329: 720 – 24.
- Filsoufi F, Castillo JG, Rahmanian PB, et al. Epidemiology of deep sternal wound infection in cardiac surgery. J Cardiothorac Vasc Anesth 2009;23:488-94.
- Koch CG, Nowicki ER, Rajeswaran J, et al. When the timing is right: antibiotic timing and infection after cardiac surgery. J Thorac Cardiovasc Surg 2012;144:931-37.

- Paul M, Raz, A, Leibovici L, et al. Sternal wound infection after coronary artery bypass graft surgery: validation of existing risk scores. J Thorac Cardiovasc Surg 2007;133:397-403.
- Lazar HL, Ketchedjian A, Haime M, et al. Topical Vancomycin in combination with perioperative antibiotics and tight glycemic control helps to eliminate sternal wound infections. J Thorac Cardiovasc Surg 2014;148:1035-40.
- Miyahara K, MatsuuraA, Takemura H, et al. Implementation of bundled interventions greatly decreases deep sternal wound infection following cardiovascular surgery. J Thorac Cardiovasc Surg 2014;148:2381-88.
- Matros E, Aranki, SF, Bayer LR, et al. Reduction in incidence of deep sternal wound infections: random or real? J Thorac Cardiovasc Surg 2010;139:680-85.
- Bardell T, Legare JF, Buth KJ, et al. ICU readmission after cardiac surgery. Eur J Cardiothorac Surg. 2003;23(3):354-359.
- Meade MO, Guyatt G, Butler R, et al. Trials comparing early vs late extubation following cardiovascular surgery. Chest. 2001:120(6 Suppl):445S-453S.
- Naughton C, Reilly N, Powroznyk A, et al. Factors determining the duration of tracheal intubation in cardiac surgery: a single-centre sequential patient audit. Eur J Anaesthesiol. 2003;20(3):225-233.
- Engel AM, McDonough S, Smith JM. Does an obese body mass index affect hospital outcomes after coronary artery bypass graft surgery? Ann Thorac Surg. 2009 Dec;88(6):1793-800.
- Brown PP, Kugelmass AD, Cohen DJ, Reynolds MR, Culler SD, Dee AD, Simon AW. The frequency and cost of complications associated with coronary artery bypass grafting surgery: results from the United States Medicare program. Ann Thorac Surg. 2008 Jun;85(6):1980-6. PubMed PMID: 18498806.
- Knapik P, Ciesla D, Borowik D, Czempik P, Knapik T. Prolonged ventilation post cardiac surgery – tips and pitfalls of the prediction game. J Cardiothorac Surg 2011;6:158.
- Hesham Z, Saleha HZ, Shawb M, et al. Outcomes and predictors of prolonged ventilation in patients undergoing elective coronary surgery. Interact Cardiovasc Thorac Surg 2012;15:51–6.
- Jacobs JP, He X, O'Brien SM, Welke KF, Filardo G, Han JM, Ferraris VA, Prager RL, Shahian DM.. Variation in Ventilation Time after Coronary Artery Bypass Grafting: An Analysis from The Society of Thoracic Surgeons Adult Cardiac Surgery Database. Ann Thorac Surg. 2013 Sep;96(3):757-62.
- Alsabbagh MM, Asmar A, Ejaz NI, Aiyer RK, Kambhampati G, Ejaz AA. Update on clinical trials for the prevention of acute kidney injury in patients undergoing cardiac surgery. Am J Surg 2013;206:86-95
- Arora P, Kolli, H, Nainani N, Nader N, Lohr J. Preventable risk factors for acute kidney injury in patients undergoing cardiac surgery. J Cardiothorac Vasc Anesth 2012; 26:687-697.
- Boldt J, Brenner T, Lehmann A, Suttner SW, Kumle B, Isgro F. Is kidney function altered by the duration of cardiopulmonary bypass? Ann Thorac Surg. 2003;75(3):906-912.
- Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. Am J Med. 1998;104(4):343-348
- Conlon PJ, Stafford-Smith M, White WD, Newman MF, King S, Winn MP, Landolfo K. Acute renal failure following cardiac surgery. Nephrol Dial Transplant. 1999;14(5):1158-1162.
- Cooper WA, O'Brien SM, Thourani VH, Guyton RA, Bridges CR, Szczech LA, Petersen R, Peterson ED. Impact of renal dysfunction on outcomes of coronary artery bypass surgery: results from the Society of Thoracic Surgeon's National Adult Cardiac Database. Circulation. 2006;113:1063-1070.
- Gallagher S, Jones DA, Lovell MJ, Hassan S, Wragg A, Kapur A, Uppal R, Yaqoob MM. The impact of acute kidney injury on midterm outcomes after coronary artery bypass graft surgery: a matched propensity score analysis. J Thorac Cardiovasc Surg 2104;147:989-995.
- Haase M, Haase-Fielitz A, Bellomo R, Devarajan P, Story D, Matalanis G, Reade MC, Bagshaw SM, Seevanayagam N, Seevanayagam S, Doolan L, Buxton B, Dragun D. Sodium bicarbonate to prevent increases in serum creatinine after cardiac surgery: a pilot double-blind, randomized trial. Crit Care Ned 2009;37:39-47.
- Hillis GS, Croal BL, Buchan KG, El-Shafei H, Gibson G, Jeffrey RR, Millar CGM, Prescott GJ, Cuthbertson BH. Renal function and outcome from coronary artery bypass grafting: impact on mortality after 2.3-year follow up. Circulation. 2006;113:1056-1062.

- Hillis LD, Smith PK, Anderson JL, Bittl JA, Bridges CR, Byrne JG, Cigarroa JE, DiSesa VJ, Hiratzka LF, Hutter AM, Jessen ME, Keeley EC, Lahey SJ, Lange RA, London MJ, Mack MJ, Patel MR, Puskas JD, Sabik JF, Selnes O, Shahian DM, Trost JC, Winniford MD. 2011 ACC/AHA guideline for coronary artery bypass graft surgery: executive summary. A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation. 2011;124:2610 -2642.
- Karthik S, Musleh G, Grayson AD, Keenan DJ, Hasan R, Pullan DM, Dihmis WC, Fabri BM. Effect of avoiding cardiopulmonary bypass in non-elective coronary artery bypass surgery: a propensity score analysis. Eur J Cardiothorac Surg. 2003;24(1):66-71.
- Kramer RS, Quinn RD, Groom RC, Braxton JH, Malenka DJ, Kellett MA, Brown JR for the Northern New England Cardiovascular Disease Study Group. Same admission cardiac catheterization and cardiac surgery: is there an increased incidence of acute kidney injury? Ann Thorac Surg 2010;90:1418-1424.
- Kuss O, von Salviati B, Borgermann J. Off-pump versus on-pump coronary artery bypass grafting: a systematic review and meta-analysis of propensity score analyses. J Thorac Cardiovasc Surg 2010;140:829-35.
- Lamy A, Devereaux PJ, Prabhakaran D, Taggart DP, Hu S, Paolasso E, Straka Z, Piegas LS, Akar AR, Jain AR, Noiseux N, Padmanabhan C, Bahamondes JC, Novick R, Vaijyanath P, Reddy S, Tao L, Olavegogeascoechea PA, Airan B, Sulling TA, Whitlock RP, Ou Y, Ng J, Chrolavicius S, Yusuf S for the CORONARY Investigators. Off-Pump or On-Pump Coronary-Artery Bypass Grafting at 30 Days. N Engl J Med 2012;366:1489-97.
- Mangano CM, Diamondstone LS, Ramsay JG, et al. Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization: the Multicenter Study of Perioperative Ischemia Research Group. Ann Intern Med. 1998;128(3):194-203.
- Medalion B, Cohen H, C, Assali A, Vaknin Assa H, Farkash A, Snir E, Sharoni E, Biderman P, Milo G, Battler A, Kornowski R, Porat E. The effect of cardiac angiography timing, contrast media dose, and preoperative renal function on acute renal failure after coronary artery bypass grafting. J Thorac Cardiovasc Surg 2010;139:1539-44.
- Ranucci M, Ballotta A, Agnelli B, Frigiola A, Mencanti L, Castelvecchio S, for the Surgical and Clinical Outcome Research (SCORE) Group. Acute kidney injury in patients undergoing cardiac surgery and coronary angiography on the same day. Ann Thorac Surg 2103;95:513-519.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Seabra VF, Alobaidi S, Balk EM, Poon AH, Jaber BL. Off-pump coronary artery bypass surgery and acute kidney injury: a meta-analysis of randomized controlled trials. Clin J Am Soc Nephrol 2010;5:1734-1744.
- Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, DeLong ER, O'Brien SM, Shewan CM, Dokholyan RS, Peterson ED. Quality Measurement in adult cardiac surgery: Part 1-conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12
- Shroyer AL, Grover FL, Hattler B, Collins JF, McDonald GO, Kozora E, Lucke JC, Baltz JH, Novitzky D, for the Veterans Affairs Randomized On/Off Bypass (ROOBY) Study Group. On-Pump versus Off-Pump Coronary-Artery Bypass Surgery. N Engl J Med 2009;361:1827-37.
- Stallwood MI, Grayson AD, Mills K, et al. Acute renal failure in coronary artery bypass surgery: independent effect of cardiopulmonary bypass. Ann Thorac Surg. 2004;77(3):968-972.
- Tang AT, Alexiou C, Hsu J, Sheppard SV, Haw MP, Ohri SK. Leukodepletion reduces renal injury in coronary revascularization: a prospective randomized study. Ann Thorac Surg. 2002;74(2):372-327; discussion 377.
- Afilalo J, Rasti M, Ohayon SM, Shimony A, Eisenberg MJ. Off-pump vs on-pump coronary bypass surgery: an updated meta-analysis and meta-regression of randomized trials. Eur Heart J. 2012; 33:1257-67
- Amory DW, Grigore A, Amory JK, et al. Neuroprotection is associated with beta-adrenergic receptor antagonists during cardiac surgery: evidence from 1,575 patients. J Cardiothorac Vasc Anesth. 2002;16(3):270-277.
- Arsenault KA, Yusus AM, Crystal E, Healey JS, Morillo CA, Nair GM et al. Interventions for preventing postoperative atrial fibrillation in patients undergoing heart surgery. Cocrane Database Syst Rev. 2013; 1:CD003611
- Bucerius J, Gummert JF, Borger MA, et al. Predictors of delirium after cardiac surgery delirium: effect of beating-heart (off-pump) surgery. J Thorac Cardiovasc Surg. 2004;127(1):57-64.

- Engelman DT, Cohn LH, Rizzo RJ. Incidence and predictors of TIAs and strokes following coronary artery bypass grafting: report and collective review. Heart Surg Forum. 1999;2(3):242-245.
- Inoue K, Luth JU, Pottkamper D, et al. Incidence and risk factors of perioperative cerebral complications: heart transplantation compared to coronary artery bypass grafting and valve surgery. J Cardiovasc Surg. 1998;39(2):201-208.
- Likosky DS, Leavitt BJ, Marrin CA, et al. Intra- and postoperative predictors of stroke after coronary artery bypass grafting. Ann Thorac Surg. 2003;76(2):428-434.
- Mangano DT. Aspirin and mortality from coronary bypass surgery. N Engl J Med. 2002; 347(17):1309-1317.
- Puskas JD, Winston AD, Wright CE, et al. Stroke after coronary artery operation: incidence, correlates, outcome and cost. Ann Thorac Surg. 2000:69(4):1053-1056.
- Brown PP, Kugelmass AD, Cohen DJ, Reynolds MR, Culler SD, Dee AD, Simon AW. The frequency and cost of complications associated with coronary artery bypass grafting surgery: results from the United States Medicare program. Ann Thorac Surg. 2008 Jun;85(6):1980-6. PubMed PMID: 18498806.
- Naylor AR. Does the risk of post-CABG stroke merit staged or synchronous reconstruction in patients with symptomatic or asymptomatic carotid disease? J Cardiovasc Surg (Torino). 2009 Feb;50(1):71-81.
- Bouchard D, Carrier M, Demers P, Cartier R, Pellerin M, Perrault LP, et al. Statin in combination with beta blocker therapy reduces postoperative stroke after coronary artery bypass graft surgery. Ann Thorac Surg. 2011:91(3) 654-9.
- Rosenberger P, Shernan SK, Loffler M, Shekar PS, Fox JA, Tuli JK, Nowak M and Eltzschig HK. The influence of epiaortic ultrasonograpy n intraoperative surgical management in 6051 cardiac surgical patients. Ann Thorac Surg. 2008; 85: 548-53.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Please see response above.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

_____

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of

evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

_____

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report
Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and

improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Evidence_Form.Surgeon_Composite_for_Adult_Cardiac_Surgery.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
N/A

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
Please see Appendix.

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
Please see Appendix.

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness
1c.2. **If Other:**

1c.3. **Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.**
Please see attached evidence form for detailed information.

1c.4. Citations for data demonstrating high priority provided in 1a.3
Please see attached evidence form for list of references.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*
N/A

1d. Composite Quality Construct and Rationale

1d.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
  - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
  - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

1d.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1d.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

Suitable for evaluating surgical performance of individual adult cardiac surgeons, the STS Individual Surgeon Composite Measure for Adult Cardiac Surgery is based on aggregate risk-adjusted morbidity and mortality for five common procedures, i.e., isolated coronary artery bypass grafting (CABG), isolated aortic valve replacement (AVR), AVR+CABG, isolated mitral valve repair or replacement (MVRR), and MVRR+CABG. Similar to other STS composite measures, this measure is based on a combination of the NQF-endorsed risk-adjusted operative mortality outcome measure and the risk-adjusted occurrence of any of five major complications. To assess overall quality, the composite comprises the following two domains:

Domain 1 – Risk-Adjusted Operative Mortality
Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Risk-Adjusted Major Morbidity
Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, coronary graft occlusion, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

This composite measure differs from the NQF-endorsed, program-level STS CABG Composite Score in that it does not include the two process measure domains (use of internal mammary artery in CABG and perioperative medications). This approach was necessary for computational reasons to efficiently combine the results from five procedures, most of which did not have comparable process measures available.

1d.3. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive

value over the component measures individually.
Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, it fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events.

In addition, with the development of this composite measure, STS addresses a number of major concerns that have previously been raised regarding surgeon-level metrics. It combines results from five of the most frequently performed cardiac surgical procedures, encompassing most of a typical adult cardiac surgeon's practice, as opposed to basing performance on just one or a few separate procedures. Furthermore, it provides a more comprehensive quality assessment and additional endpoints, as it includes risk-adjusted mortality and the risk-adjusted occurrence of any of five major complications. This measure will be useful to surgeons in identifying potential areas for improvement, and it has numerous advantages compared with existing surgeon metrics if used for accountability purposes.

1d.4. **Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.**
The mortality domain corresponds to a single measure, while the study endpoint for the morbidity domain combines multiple measures and thus is a composite endpoint.

Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviation across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular, Surgery, Surgery : Cardiac Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Complications, Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
Data collection form: http://www.sts.org/sites/default/files/documents/STSAdultCVDataCollectionForm2_73_Annotated.pdf; Data specifications:
http://www.sts.org/sites/default/files/documents/word/STSAdultCVDataSpecificationsV2_73%20with%20correction.pdf;

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
See Appendix

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
See Appendix

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See Appendix

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
See Appendix

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See Appendix

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model

If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See Appendix

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
See Appendix

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please see discussion under section S.4 and attached manuscripts.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.4% of records for operative mortality and 0.3% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses

with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

The overall frequency of missing data was 0.4% for operative mortality and 0.3% for major complications. The median surgeon-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications. The percent of surgeons with >10% missing data was 1.0% for mortality and 1.0% for major complications. As a sensitivity analysis, we recalculated surgeon-specific mortality and complication rates after excluding records with missing data from the denominator. As shown in section 2b7.2. of the testing attachment, there was high (>0.99) correlation between surgeon-specific rates calculated with missing data excluded versus imputed.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database – Version 2.73; STS Adult Cardiac Surgery Database Version 2.81 went live on July 1, 2014, but there were not sufficient data available in version 2.81 to develop this composite measure.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Individual

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Please see section S.4 and Appendix.

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Testing_Form.Surgeon_Composite_for_Adult_Cardiac_Surgery-636008079864847729.doc

Measure Number (*if previously endorsed*)**: Click here to enter NQF number**
Composite Measure Title: Individual Surgeon Composite Measure for Adult Cardiac Surgery
Date of Submission: ~~6/5/2016~~6/5/2016
Composite Construction:
☒Two or more individual performance measure scores combined into one score
☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)
☒ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.**
- **For composites with outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitions** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2), validity (2b2-2b6), and composites (2d) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that

results are distorted without the exclusion; [12]

AND

If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. 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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2d1. 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

2d2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(*if not conducted or results not adequate, justification must be submitted and accepted)*

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. If different data sources are used for different components in the composite, indicate the component after the checkbox.*)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record

☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: **Click here to describe**	☐ other: **Click here to describe**

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

STS Adult Cardiac Surgery Database Version 2.73

1.3. What are the dates of the data used in testing?

July 2011 – June 2014

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☐ group/practice	☐ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: **Click here to describe**	☐ other: **Click here to describe**

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The measure was developed and tested using STS data from 2286 surgeons for patients undergoing cardiac surgery during July 2011 – June 2014. Only surgeons with at least 10 eligible records during this period were included in the hierarchical model for estimating composite scores. The table below summarizes the distribution of surgeon-specific denominators (number of eligible patients) and surgeon-specific mortality and morbidity rates.

Stat	**N (Denominator)**	**% Mortality**	**% Morbidity**
N	2286	2286	2286
Mean	272	2.5	14.8
STD	166	1.8	6.7

IQR	208	2.0	7.7
0%	10	0.0	0.0
10%	79	0.7	7.7
20%	132	1.1	9.5
30%	170	1.5	10.9
40%	209	1.9	12.3
50%	250	2.2	13.8
60%	293	2.6	15.3
70%	336	3.1	17.0
80%	394	3.7	19.2
90%	478	4.8	23.0
100%	1435	15.4	64.8

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

The study cohort included 621,489 patient operations performed by 2,286 surgeons during July 2011 – June 2014. The number of patients and their unadjusted outcomes are summarized by procedure type in the following table.

	Total	Mortality		Morbidity	
Procedure Group	No. (%)	No.	Rate, %	No.	Rate, %
Isolated CABG	417,261 (67.1)	8,295	2.0	51,281	12.3
Isolated AVR	84,751 (13.6)	2,059	2.4	11,458	13.5
Isolated MVR	14,948 (2.4)	539	3.6	2,905	19.4
AVR + CABG	53,081 (8.5)	2,124	4.0	10,801	20.3
MVR + CABG	6,547 (1.1)	474	7.2	2,125	32.5
Isolated MV repair	30,347 (4.9)	339	1.1	2,953	9.7
MV repair + CABG	14,554 (2.3)	635	4.4	3,694	25.4
Total	621,489 (100.0)	14,465	2.3	85,217	13.7

AVR = aortic valve replacement; CABG = coronary artery bypass grafting; MV = mitral valve; MVR = mitral valve replacement.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

The methodology was developed and tested using data from 621,489 patients operations performed by 2,286 surgeons. To ensure adequate statistical precision, STS plans to report composite scores only for surgeons with at least 100 eligible cases during the 3-year measurement window. Thus, some of the analyses in this submission are limited to surgeons with at least 100 eligible cases.

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

Note: *Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. This quantity cannot be calculated directly because the "true" composite measure values are unknown, but may be estimated, as described below.

Calculation Details

Let θ_j denote the true unknown composite measure value for the j-th of J surgeons. Before estimating reliability, the numeric value of θ_j was estimated for each surgeon under the assumed hierarchical model. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. For each *j,* we randomly generated a large number (*N*) of possible numeric values of θ_j by sampling from the Bayesian posterior probability distribution of θ_j via MCMC sampling. Let $\theta_j^{(i)}$ denote the i-th of these *N* randomly sampled numerical values for the j-th surgeon.

2. For each j, the posterior mean $\hat{\theta}_j$ of θ_j was calculated as the arithmetic average of the randomly sampled values $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$; in other words $\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N} \theta_j^{(i)}$.

Our reliability measure was defined as the squared correlation between the set of hospital-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and the corresponding unknown true values $\theta_1, \ldots, \theta_J$. Let ρ^2 denote the unknown true squared correlation of interest and let $\hat{\rho}^2$ denote an estimate of this quantity. The estimate was calculated as

$$\hat{\rho}^2 = \frac{1}{N}\sum_{i=1}^{N} \rho_{(i)}^2$$

where

$$\rho_{(i)}^2 = \frac{\left[\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)\left(\hat{\theta}_j - \bar{\theta}\right)\right]^2}{\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)^2 \sum_{j=1}^{J}\left(\hat{\theta}_j - \bar{\theta}\right)^2}, \quad \bar{\theta} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N} \theta_j^{(i)} \quad \text{and} \quad \bar{\theta}^{(i)} = \frac{1}{J}\sum_{j=1}^{J} \theta_j^{(i)}.$$

A 95% Bayesian probability interval for ρ^2 was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers $\rho_{(1)}^2, \ldots, \rho_{(N)}^2$.

2a2.3. What were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

The estimated reliability of the composite measure using 3 years of data in surgeons with at least 100 total cases was 0.81 (95% CrI, 0.79 to 0.82), as outlined in the Table below. For comparison, reliability of the STS isolated CABG composite score was 0.77 (95% CrI, 0.74 to 0.80) using 1 year of data in 2013. Using 3 years of data from 2011 to 2013, the reliability of the STS AVR composite measure was 0.52 (95% CrI, 0.47 to 0.57), and the AVR+CABG measure was 0.50 (95% CrI, 0.45 to 0.54).

Threshold Number of Index Cases Over 3 Years	Surgeons Included (No.)	Patients Included (No.)	Reliability $\hat{\rho}^2$ (95% PrI)
10	2,286	621,489	0.77 (0.75, 0.79)
25	2,234	620,586	0.78 (0.76, 0.80)
36	2,205	619,691	0.79 (0.77, 0.80)
50	2,165	617,976	0.79 (0.78, 0.81)
100	1,976	603,594	0.81 (0.79, 0.82)
150	1,737	573,491	0.81 (0.80, 0.83)
200	1,432	520,724	0.82 (0.81, 0.84)

[a] Number of surgeons and patients included for each threshold.

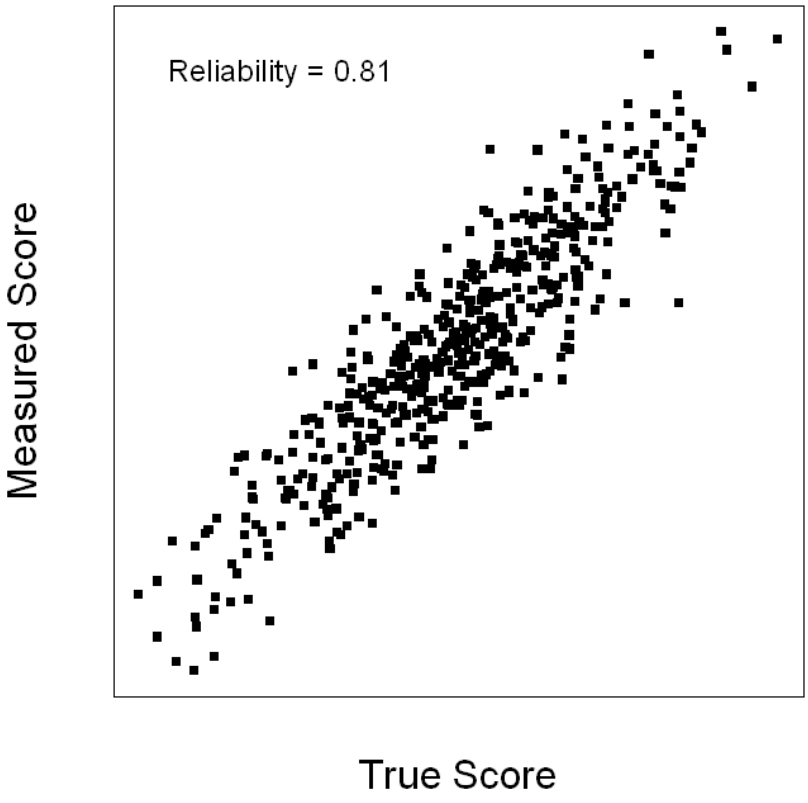
$\hat{\rho}^2$ is the estimated squared correlation between the set of surgeon-specific estimates of composite performance measure values and their corresponding unknown true values (used as the measure of reliability in this study).

PrI = probability interval.

Based in part on these results, we selected a threshold of 100 cases over 3 years, as a minimum threshold for receiving a surgeon-specific composite score. This resulted in a reliability of 0.81 but reduced the number of surgeons eligible to receive a score from 2,286 to 1,976. A higher volume threshold would have yielded even higher reliability but at the cost of further reducing the number of surgeons eligible to receive a score.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

To interpret the results, we created a figure illustrating the accuracy of the measured scores when the true reliability = 0.81. Because the true score for the composite measure is unknown, we used simulated data with formula Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2, \ldots, 1976$ indicates the 1,976 surgeons and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure has a reliability of 0.81.



2b2. VALIDITY TESTING

Note: *Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b2.1. What level of validity testing was conducted?

☒ **Composite performance measure score**

- ☐ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)
- ☐ **Systematic assessment of content validity**

☐ **Validity testing for component measures** *(check all that apply)*

Note: *applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

- ☐ **Endorsed (or submitted) as individual performance measures**
- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Empirical validity testing of the component measure score(s)**
- ☐ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

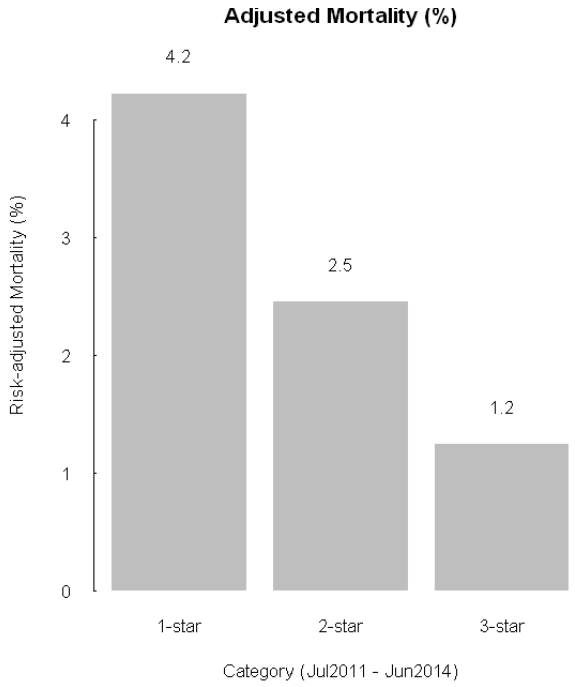
2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

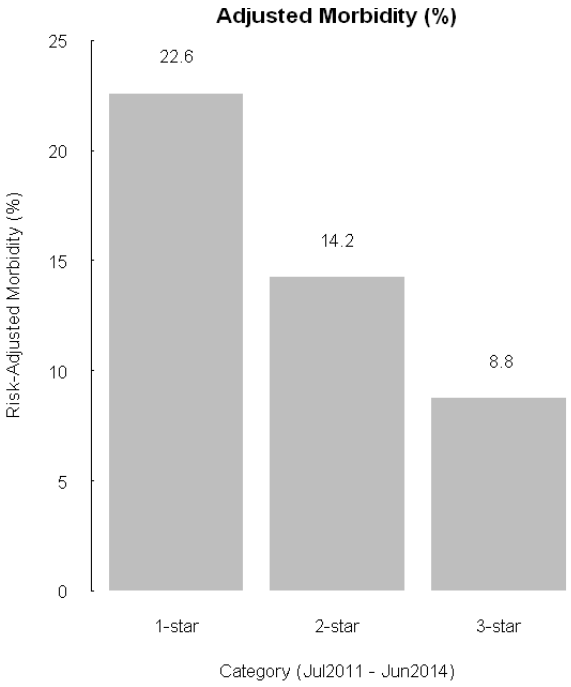
The tests on validity used the concept of performance categories to be more formally introduced in 2b5: Surgeons were labeled as having higher-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely above the overall STS average composite score. Surgeons were labeled as having lower-than-expected performance if the 98% credible interval surrounding a surgeon's composite score fell entirely below the overall STS average composite score. Surgeons were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).

We compared risk-adjusted mortality and morbidity rates across the three performance groups. The measure has good face value if the three groups have different proportions as expected.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Compared to surgeons receiving 1 star, those with 3 stars had lower risk-adjusted mortality (1.2% vs. 4.2%) and lower risk-adjusted morbidity (8.8% vs. 22.6%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS surgeons deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.





2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance, and observed differences in morbidity and mortality rates correspond appropriately with the changes in

performance categories. These results support the validity of the composite measure as a quality measure for cardiac surgery.

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

N/A

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

N/A

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☒ **Endorsed (or submitted) as individual performance measures**
☐ **No risk adjustment or stratification**
☒ **Statistical risk model**
☐ **Stratification by risk categories**
☐ **Other, Click here to enter description**

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

Surgeon-specific risk-adjusted operative mortality and major complication rates were estimated using a bivariate random-effects logistic regression model [1]. To adjust for case mix, each patient's risk score for operative mortality and his or her corresponding risk score for major complications were first calculated, using existing and modified STS risk

models as described below. The goal of calculating a risk score was to reduce the number of covariates in the hierarchical model by summarizing the predictive information from a large number of baseline covariates into a single number. Adjustment for each covariate individually in the hierarchical model would be theoretically preferable but is impractical due to the large number of records and covariates and the computationally intensive nature of Bayesian hierarchical model estimation.

To study the consequences of using an overall risk score for each patient instead of individual covariates in our models, sensitivity analyses were performed in which each surgeon's risk-adjusted mortality and complication rates were estimated in models that adjusted for 41 and 47 individual patient covariates, respectively. These estimates were compared with those derived from models adjusting for a single composite risk score. To make this analysis computationally manageable, model variables were estimated by maximum likelihood (i.e., empirical Bayes) instead of performing a fully Bayesian analysis. To further simplify this sensitivity analysis, mortality and complication rates were estimated in separate models, not simultaneously in a single model, and the cohort was restricted to isolated CABG. For each end point (operative mortality and major complications) we calculated each surgeon's risk-standardized rate of the end point using each model and compared the results.

After sensitivity analyses demonstrated the validity of this risk score approach, the operative mortality risk score (predicted risk of death) was then used as a covariate in the hierarchical model for operative mortality, and the major complication risk score (predicted risk of major morbidity) was used as a covariate in the hierarchical model for major complications. To reduce potential bias, the hierarchical model included both individual patient-level risk scores and the average value of these patient-level risk scores calculated separately for each surgeon [1].

For patients undergoing isolated CABG, isolated AVR, or AVR + CABG, risk scores were calculated according to the published STS 2008 mortality and major complications models for isolated CABG, isolated valve, or valve + CABG [2-4]. To ensure high calibration for the current study cohort, coefficients of each model were re-estimated using the current 3-year study sample and current end point definitions. Risk scores for patients undergoing a mitral operation without CABG were calculated using a modified version of the published STS 2008 mortality and major complications models for isolated valve procedures [3]. These modifications allowed inclusion of patients undergoing tricuspid repair, an increasingly common adjunct, urgent and emergency procedures, all arrhythmia ablation procedures for atrial fibrillation, atrial septal defect and patent foramen ovale closures, and active and treated endocarditis. Also included are more granular classifications and adjustment for the degree of tricuspid regurgitation (less than moderate, moderate, and severe).

Coefficients of the modified models were estimated using the current 3-year study cohort and end point definitions. Risk scores for patients undergoing a mitral operation with concomitant CABG were calculated using a similarly modified version of the published STS 2008 mortality and major complications models for valve + CABG operations [4], also with re-estimated coefficients.

References

- Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, Fazzalari FL, Filardo G, Normand SL, Furnary AP, Magee MJ, Rankin JS, Welke KF, Han J, O'Brien SM. The Society of Thoracic Surgeons Composite Measure of Individual Surgeon Performance for Adult Cardiac Surgery: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2015;100:1315-25.
- Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.

- O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
- Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009 Jul;88(1 Suppl):S43-62.

2b4.4. What were the statistical results of the analyses used to select risk factors?

For isolated CABG, isolated AVR, and AVR+ CABG, risk adjustment is based on existing published risk models [1-3]. Methods and results for selecting covariates may be found in those publications.

For mitral valve repair and replacement (MVRR) and for MVRR + CABG, the published models were modified to account for the inclusion of patients undergoing tricuspid repair. Estimated odds ratios from these modified STS 2008 models are summarized in the tables below.

Odds ratios for the modified MVRR model

	Morbidity		Mortality	
Effect	OR (95% CI)	P-value	OR (95% CI)	P-value
Effects that do not interact with MV repair/replacements				
Preoperative atrial fibrillation	1.14 (1.08, 1.20)	<.0001	1.22 (1.09, 1.36)	0.0005
Race (v. others)				
Black	1.25 (1.15, 1.35)	<.0001	NA	.
Hispanic	1.23 (1.10, 1.39)	0.0003	NA	.
CVD (v. no)				
CVD with CVA	1.17 (1.08, 1.27)	0.0002	NA	.
CVD without CVA	0.97 (0.88, 1.08)	0.6072	NA	.
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.06 (1.00, 1.12)	0.0681	NA	.
Pre-op IABP or inotrope	2.20 (1.93, 2.52)	<.0001	1.35 (1.10, 1.66)	0.0040
Hypertension	1.13 (1.06, 1.20)	<.0001	1.11 (0.98, 1.27)	0.1115
Immunosuppressive treatment	1.26 (1.13, 1.41)	<.0001	1.60 (1.33, 1.93)	<.0001
Peripheral vascular disease	1.24 (1.14, 1.35)	<.0001	1.31 (1.13, 1.52)	0.0004
Aortic stenosis	1.12 (0.99, 1.27)	0.0714	NA	.
MI <21 days	1.40 (1.19, 1.63)	<.0001	1.77 (1.41, 2.21)	<.0001
Shock	2.52 (2.08, 3.06)	<.0001	1.84 (1.38, 2.47)	<.0001
Number of previous operations (v. 0)				
1 previous operation	1.35 (1.21, 1.50)	<.0001	1.69 (1.34, 2.15)	<.0001
2 or more previous operations	1.65 (1.39, 1.95)	<.0001	2.12 (1.53, 2.94)	<.0001
Urgent status (v. elective)	1.32 (1.23, 1.41)	<.0001	1.27 (1.12, 1.44)	0.0003
Active infections endocarditis	1.80 (1.63, 1.99)	<.0001	1.86 (1.54, 2.23)	<.0001
Treated infections endocarditis	1.07 (0.96, 1.19)	0.2471	0.96 (0.76, 1.22)	0.7427
Ejection fraction per 10-unit decrease	1.11 (1.07, 1.15)	<.0001	1.17 (1.09, 1.25)	<.0001
Creatinine per 1 unit increase	1.62 (1.53, 1.71)	<.0001	1.48 (1.36, 1.62)	<.0001
Body surface area, m^2				
1.6 v. 2.0 in male	1.26 (1.11, 1.44)	0.0004	1.64 (1.29, 2.09)	<.0001
1.8 v. 2.0 in male	1.05 (1.00, 1.10)	0.0418	1.19 (1.08, 1.30)	0.0002
2.2 v. 2.0 in male	1.09 (1.05, 1.13)	<.0001	0.99 (0.91, 1.06)	0.6966
1.6 v. 1.8 in female	1.07 (1.03, 1.12)	0.0017	1.25 (1.16, 1.35)	<.0001
2.0 v. 1.8 in female	1.04 (1.01, 1.08)	0.0074	0.98 (0.92, 1.04)	0.4325
2.2 v. 1.8 in female	1.21 (1.12, 1.31)	<.0001	1.17 (1.00, 1.37)	0.0500

Time trend (half year increase)	0.96 (0.95, 0.98)	<.0001	1.01 (0.98, 1.04)	0.6571
Left main disease	NA	.	1.11 (0.81, 1.53)	0.5091
Unstable angina (no MI < 8days)	NA	.	1.21 (0.94, 1.55)	0.1420
Mitral stenosis	NA	.	1.05 (0.91, 1.20)	0.5060
Moderate tricuspid insufficiency (v. no-mild)	1.13 (1.05, 1.20)	0.0003	1.17 (1.02, 1.33)	0.0243
Severe tricuspid insufficiency (v. no-mild)	1.23 (1.12, 1.35)	<.0001	1.46 (1.22, 1.75)	<.0001
Mitral valve repair (v. replacement)	0.56 (0.50, 0.62)	<.0001	0.40 (0.31, 0.53)	<.0001
Tricuspid valve repair (v. none)	1.36 (1.24, 1.49)	<.0001	0.99 (0.84, 1.18)	0.9474
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs				
In MV replacements				
Age				
60 v. 50 (no reoperations, non-emergent)	1.22 (1.17, 1.27)	<.0001	1.53 (1.40, 1.67)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.49 (1.37, 1.62)	<.0001	2.34 (1.95, 2.81)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.80 (1.61, 2.01)	<.0001	3.69 (2.95, 4.63)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.08 (1.00, 1.17)	0.0517	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.53 (1.38, 1.69)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)				
Insulin diabetes	1.41 (1.27, 1.57)	<.0001	1.48 (1.25, 1.75)	<.0001
Non-insulin diabetes	1.10 (1.02, 1.20)	0.0174	1.14 (1.00, 1.31)	0.0587
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.15 (1.11, 1.18)	<.0001	1.20 (1.13, 1.28)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	2.59 (2.12, 3.15)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.32 (1.13, 1.53)	0.0004
Status (v. elective)				
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	2.38 (1.61, 3.49)	<.0001
Emergent+resuscitation/emergent salvage	2.83 (1.63, 4.89)	0.0002	5.91 (3.18, 10.98)	<.0001
In MV repairs				
Age				
60 v. 50 (no reoperations, non-emergent)	1.27 (1.21, 1.32)	<.0001	1.76 (1.57, 1.97)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.60 (1.47, 1.75)	<.0001	3.09 (2.46, 3.88)	<.0001
80 v. 50 (no reoperations, non-emergent)	2.00 (1.78, 2.25)	<.0001	5.61 (4.19, 7.50)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.17 (1.07, 1.28)	0.0007	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.55 (1.36, 1.76)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)				
Non-insulin diabetes	1.10 (0.99, 1.21)	0.0690	1.14 (1.00, 1.31)	0.0587
Insulin diabetes	1.44 (1.24, 1.66)	<.0001	1.48 (1.25, 1.75)	<.0001
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.15 (1.11, 1.18)	<.0001	1.26 (1.16, 1.38)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	3.68 (2.44, 5.54)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.14 (0.93, 1.40)	0.2108
Status (v. elective)				
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	3.83 (1.87, 7.86)	0.0003
Emergent+resuscitation/Emergent Salvage	2.83 (1.63, 4.89)	0.0002	2.47 (0.10, 59.60)	0.5785

CHF = congestive heart failure; CVA = cerebrovascular accident (stroke); CVD = cardiovascular disease; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; NA = variable not used in model and estimate not available; NYHA = New York Heart Association.

Odds ratios for the modified MVRR + CABG model

	Morbidity		Mortality	
Effect	OR (95% CI)	P-value	OR (95% CI)	P-value
Effects that do not interact with MV repair/replacements				
Preoperative atrial fibrillation	1.09 (1.02, 1.17)	0.0125	1.04 (0.92, 1.18)	0.4926
Race (v. others)				
Black	1.21 (1.08, 1.35)	0.0007	NA	.
Hispanic	1.16 (1.00, 1.35)	0.0529	NA	.
CVD (v. no)				
CVD with CVA	1.21 (1.10, 1.33)	0.0001	1.01 (0.86, 1.19)	0.9223
CVD without CVA	1.09 (0.98, 1.21)	0.1264	NA	.
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.16 (1.11, 1.21)	<.0001	1.16 (1.08, 1.26)	<.0001
Pre-op IABP or inotrope	2.21 (1.98, 2.47)	<.0001	1.43 (1.22, 1.69)	<.0001
Hypertension	1.11 (1.02, 1.20)	0.0189	NA	.
Immunosuppressive treatment	1.17 (1.02, 1.34)	0.0264	1.29 (1.02, 1.63)	0.0303
Peripheral vascular disease	1.08 (1.00, 1.17)	0.0536	1.28 (1.11, 1.48)	0.0007
MI (v. no recent MI)				
1-21 days	1.32 (1.23, 1.42)	<.0001	1.30 (1.13, 1.50)	0.0002
<=24 hrs	1.48 (1.16, 1.89)	0.0015	1.76 (1.28, 2.40)	0.0004
Number of previous operations (v. 0)				
1 previous operation	1.45 (1.15, 1.83)	0.0017	2.79 (1.88, 4.14)	<.0001
2 or more previous operations	1.50 (1.00, 2.24)	0.0485	2.68 (1.41, 5.06)	0.0025
Diabetes (v. no)				
Non-insulin diabetes	1.22 (1.12, 1.32)	<.0001	1.35 (1.17, 1.57)	<.0001
Insulin diabetes	1.08 (1.01, 1.16)	0.0233	1.10 (0.97, 1.24)	0.1565
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.10 (1.07, 1.14)	<.0001	1.16 (1.10, 1.22)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	2.17 (1.88, 2.50)	<.0001	2.66 (2.19, 3.23)	<.0001
Creatinine per 1 unit increase	1.62 (1.51, 1.73)	<.0001	1.46 (1.33, 1.61)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.20 (1.11, 1.29)	<.0001	1.39 (1.21, 1.59)	<.0001
Status (v. elective)				
Urgent	1.26 (1.18, 1.36)	<.0001	1.09 (0.96, 1.24)	0.1821
Emergent - no resuscitation	2.53 (1.75, 3.65)	<.0001	1.74 (1.12, 2.73)	0.0148
Emergent+resuscitation/Emergent Salvage	1.90 (1.07, 3.38)	0.0292	5.13 (2.83, 9.31)	<.0001
Active infections endocarditis	1.48 (1.20, 1.83)	0.0003	1.63 (1.18, 2.24)	0.0027
Treated infections endocarditis	0.91 (0.72, 1.16)	0.4538	0.57 (0.33, 0.97)	0.0393
Body surface area, m^2				
1.6 v. 2.0 in male	1.16 (1.01, 1.34)	0.0354	1.32 (1.02, 1.72)	0.0354
1.8 v. 2.0 in male	1.02 (0.97, 1.08)	0.4400	1.07 (0.97, 1.17)	0.1703
2.2 v. 2.0 in male	1.09 (1.05, 1.14)	<.0001	1.08 (1.01, 1.16)	0.0234
1.6 v. 1.8 in female	1.12 (1.06, 1.18)	0.0002	1.24 (1.12, 1.36)	<.0001
2.0 v. 1.8 in female	1.06 (1.00, 1.12)	0.0360	1.03 (0.94, 1.12)	0.5595
2.2 v. 1.8 in female	1.33 (1.15, 1.54)	0.0002	1.34 (1.06, 1.68)	0.0133
Time trend (half year increase)	0.98 (0.96, 1.00)	0.0541	1.03 (1.00, 1.06)	0.0440
Left main disease	NA	.	1.09 (0.96, 1.24)	0.1778
Unstable angina (no MI < 8days)	NA	.	1.01 (0.87, 1.17)	0.9382
Mitral stenosis	NA	.	1.21 (1.01, 1.46)	0.0399
Mitral insufficiency (>= moderate)	0.95 (0.86, 1.05)	0.3396	NA	.
Moderate tricuspid insufficiency (v. no-mild)	1.10 (1.02, 1.20)	0.0189	1.10 (0.96, 1.26)	0.1618

Severe tricuspid insufficiency (v. no-mild)	1.12 (0.98, 1.29)	0.1051	1.12 (0.89, 1.41)	0.3448
Mitral valve repair (v. replacement)	0.69 (0.59, 0.81)	<.0001	0.81 (0.59, 1.10)	0.1784
Tricuspid valve repair (v. none)	1.33 (1.19, 1.49)	<.0001	1.04 (0.85, 1.27)	0.7010
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs				
In MV replacements + CABG				
Age				
60 v. 50 (no reoperations, non-emergent)	1.16 (1.09, 1.23)	<.0001	1.70 (1.51, 1.91)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.35 (1.20, 1.52)	<.0001	2.88 (2.28, 3.64)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.57 (1.34, 1.84)	<.0001	4.84 (3.62, 6.49)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.15 (1.04, 1.28)	0.0063	1.14 (0.94, 1.37)	0.1794
CHF NYHA IV	1.36 (1.18, 1.55)	<.0001	1.49 (1.21, 1.83)	0.0002
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.04 (0.96, 1.14)	0.3436
Shock	2.07 (1.59, 2.69)	<.0001	1.89 (1.49, 2.39)	<.0001
In MV repairs + CABG				
Age				
60 v. 50 (no reoperations, non-emergent)	1.16 (1.10, 1.21)	<.0001	1.45 (1.31, 1.61)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.34 (1.21, 1.47)	<.0001	2.11 (1.72, 2.60)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.55 (1.36, 1.76)	<.0001	3.04 (2.35, 3.92)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.15 (1.05, 1.27)	0.0027	1.27 (1.06, 1.51)	0.0087
CHF NYHA IV	1.32 (1.18, 1.49)	<.0001	1.40 (1.14, 1.73)	0.0016
Shock	1.97 (1.56, 2.47)	<.0001	1.89 (1.49, 2.39)	<.0001
Ejection fraction per 10-unit decrease	1.12 (1.09, 1.16)	<.0001	1.13 (1.06, 1.21)	0.0002

CHF = congestive heart failure; CVA = cerebrovascular accident (stroke); CVD = cardiovascular disease; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; NA = variable not used in model and estimate not available; NYHA = New York Heart Association.

References

1. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
2. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
3. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009 Jul;88(1 Suppl):S43-62.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

For isolated CABG, isolated AVR, and AVR+ CABG, risk adjustment is based on existing published risk models [1-3]. Model performance metrics may be found in those publications.

For MVRR, modifications to the existing STS models were assessed using data from 62,118 patients undergoing MVRR during July 2011 – June 2014.

For MVRR + CABG, modifications to the existing STS models were assessed using data from 26,355 patients undergoing MVRR + CABG during July 2011 – June 2014.

Discrimination

Discrimination results are presented for the modified MVRR and MVRR + CABG models. To gauge discrimination, we calculated the c-statistics of both models. Bootstrapping was used to estimate and adjust for the "optimism" from estimating and evaluating the model on the same sample [4].

Calibration

Calibration results are presented for the modified MVRR and MVRR + CABG models. The model fit was evaluated using 5-fold cross validation. The entire sample was randomly split into five equal-sized groups. The calibration plot was created by following these steps:

1. One of the five groups was used as the testing sample
2. The other four groups were combined into the training sample
3. The revised model was estimated using the training sample
4. The expected probability of experience the event in the testing sample was calculated using the model estimated in step 3.
5. The expected probability (from step 4) and observed event rates were then compared in the testing sample and the calibration plot was created.

The above five steps were repeated five times so that each group was used as the testing sample once. In the end, we had five calibration plots for each model.

References

1. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1--coronary artery bypass grafting surgery. Ann Thorac Surg. 2009 Jul;88(1 Suppl):S2-22.
2. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
3. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3--valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009 Jul;88(1 Suppl):S43-62.
4. Harrell, F. E., Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine 15 (1996): 361-387.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

Modified MVRR model

The bootstrap-adjusted estimated C-statistic was 0.746 for the morbidity model and 0.807 for the mortality model. These numbers were comparable to the STS 2008 models when evaluated using the same sample (0.745 and 0.807 for morbidity and mortality endpoints, respectively.)

Modified MVRR + CABG model

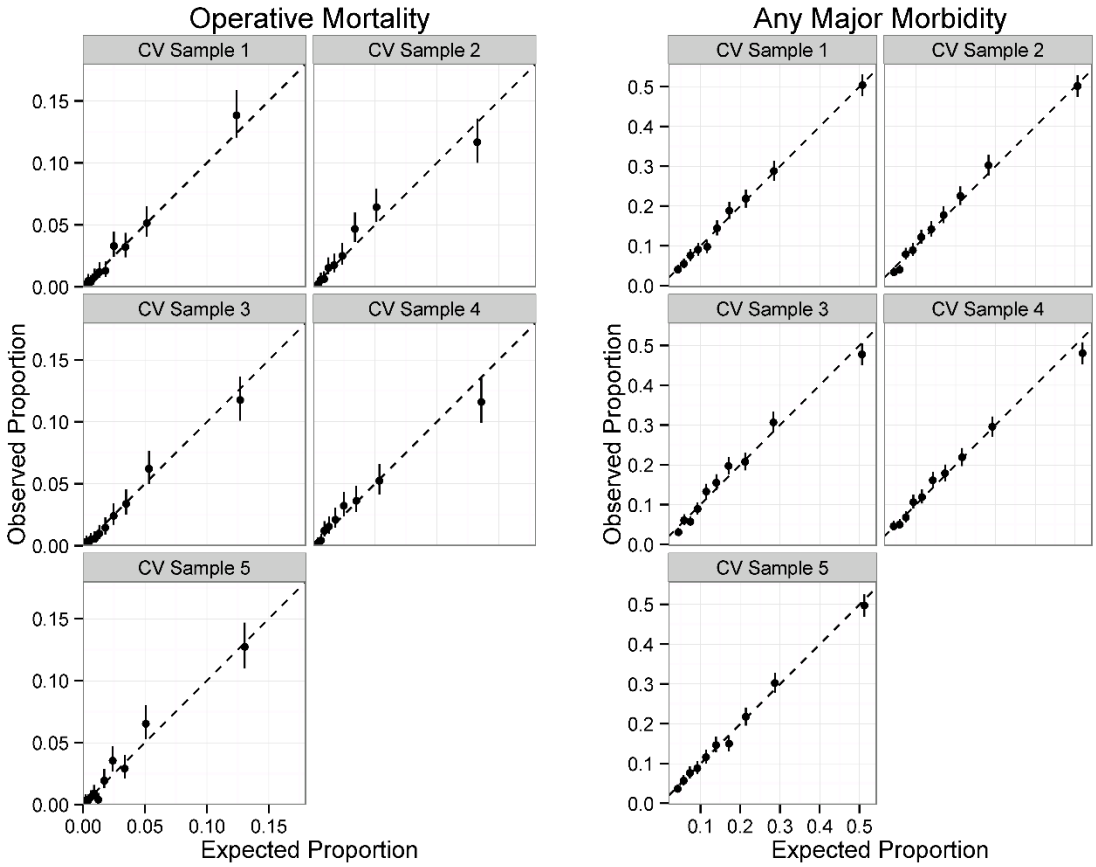
The bootstrap-adjusted C statistic was 0.708 for the morbidity model and 0.738 for the mortality model. These numbers were comparable to the STS 2008 models when evaluated using the same sample (0.707 and 0.738 for morbidity and mortality endpoints, respectively.)

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A. The Hosmer-Lemeshow statistic was not calculated.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

Modified MVRR model



Modified MVRR + CABG model

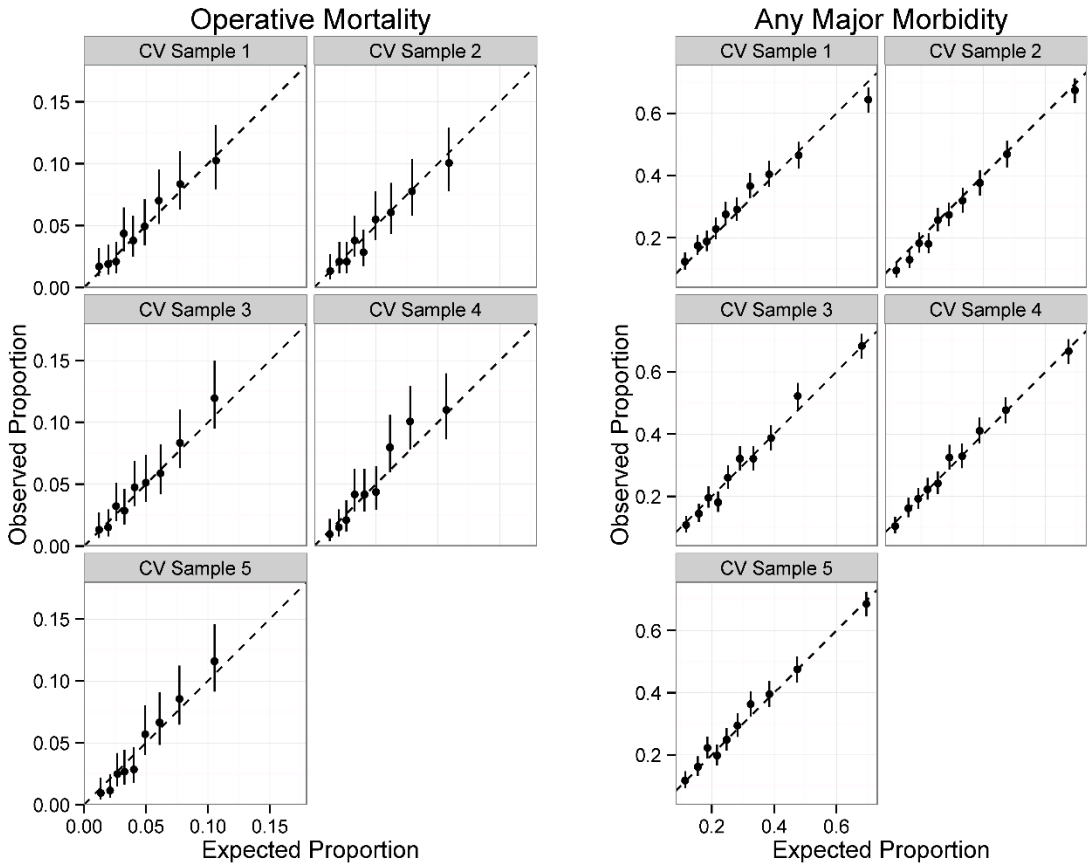


Figure. Plots of observed versus expected in cross validation samples, operative mortality

2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The results demonstrated that the STS cardiac surgery risk models are well calibrated and have good discrimination power. They are suitable for controlling differences in case-mix between centers.

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

Note: *Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The degree of uncertainty surrounding an STS surgeon's composite measure estimate is indicated by calculating 98% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an individual STS surgeon are reported along with a comparison to various benchmarks based on the national sample. Benchmarks include the overall average STS composite score and several percentiles (minimum, 10^{th}, 25^{th}, 75^{th}, 90^{th}, maximum). In addition, the composite measure result is converted into categories labeled as 1, 2 and 3 stars. An STS surgeon receives 2 stars if the Bayesian credible interval surrounding his/her composite score overlaps the overall STS average. This rating implies that the STS surgeon's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the surgeon receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the surgeon receives 1 star (lower-than-expected performance).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Among surgeons with at least 100 cases over 3 years, around 71% of surgeons received 2 stars, and the remaining surgeons received either 1 or 3 stars.

Performance categories July 2011 – June 2014

	All Surgeons	**Surgeons N≥ 100**
Category	**Number of Surgeons, %**	**Number of Surgeons, %**
1-star	207, 9.1%	189, 9.6%
2-star	1701, 74.4%	1413, 71.5%
3-star	378, 16.5%	374, 18.9%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The Bayesian methodology allows direct probability interpretation of the results. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of surgeons across performance categories.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.* ***If only one set of specifications for each component, this section can be skipped.***

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted?*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

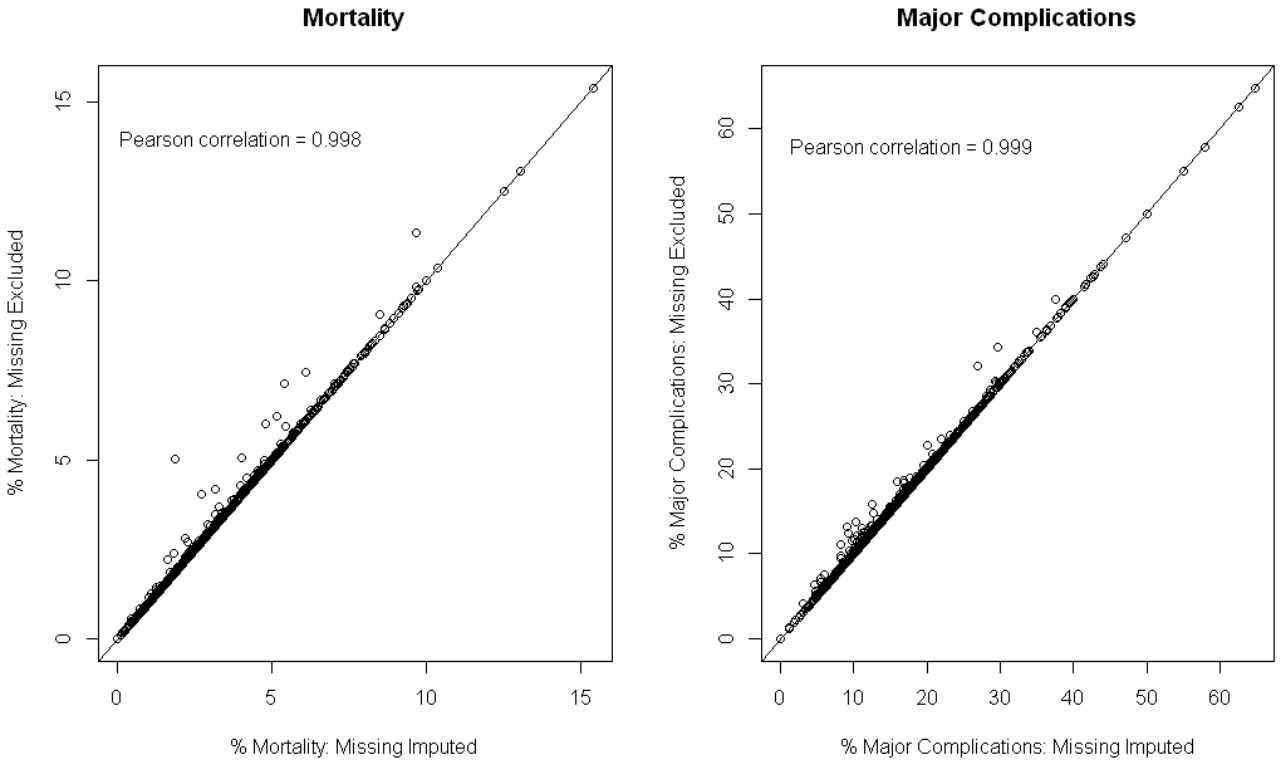
Missing data for risk model covariates were extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.4% of records for operative mortality and 0.3% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The overall frequency of missing data was 0.4% for operative mortality and 0.3% for major complications. The median surgeon-specific frequency of missing data was 0% (range 0% to 65%) for mortality and 0% (range 0% to 40%) for major complications. The percent of surgeons with >10% missing data was 1.0% for mortality and 1.0% for major complications. As a sensitivity analysis, we re-calculated surgeon-specific mortality and complication rates after

excluding records with missing data from the denominator. As shown in the figure below, there was high (>0.99) correlation between surgeon-specific rates calculated with missing data excluded versus imputed.



2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

These results suggest that our handling of missing outcome data is unlikely to impact performance results for the vast majority of surgeons.

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To verify that each domain contributes statistical information but does not dominate the composite, we calculated the correlations between each domain-specific estimate and the overall composite score. These analyses were performed using data from July 2011 – June 2014.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

Pearson Correlation With Overall Composite	
Mortality	**Morbidity**
0.73	0.92

The Pearson correlations were 0.73 for mortality versus overall composite measure and 0.92 for morbidity domain score versus overall score.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

Although risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate. Both domains contribute statistical information.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

The overall composite score was calculated as a weighted sum of (1 minus risk-adjusted mortality rate) and (1 minus risk-adjusted major morbidity rate). Mortality and morbidity rates were weighted inversely by their respective standard deviations across surgeons. This procedure is equivalent to first rescaling mortality and morbidity rates by their respective standard deviation across surgeons and then assigning equal weighting to the rescaled mortality rate and rescaled morbidity rate. Standard deviations derived from the data were used to define the final composite measure as 0.81 × (1 minus risk-standardized mortality rate) + 0.19 × (1 minus risk-standardized complication rate).

This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains. To facilitate the assessment, we calculated for a 1 percentage point change in mortality, what percentage point change in morbidity would be needed to achieve the same impact on the composite measure.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.81 and 0.19, respectively. An implication of this weighting is that a 1 percentage point change in a surgeon's risk-adjusted mortality rate has the same impact as a 4.3 percentage point change in the site's risk-adjusted morbidity rate.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. **Data Elements Generated as Byproduct of Care Processes.**

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)

If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*

Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**

The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**

No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**

IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.

The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

Data Collection:

There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. **Current and Planned Use**

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. **For each CURRENT use, checked above, provide:**

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. **If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons?** (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new composite measure, which was developed in 2014 and published in 2015. STS is currently in the planning stages of making results available to individual surgeons and anticipate distribution in late 2016 or early 2017.

4a.3. **If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement.** (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data*

aggregation and reporting.)
STS plans to roll out public reporting of this composite measure within the next several years.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. **Progress on Improvement. (Not required for initial endorsement unless available.)**
Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Data are provided in 1b.2 and 1b.4 as required.

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**
N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**
We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2014, 10% of participants were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures
Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.
Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**
0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0116 : Anti-Platelet Medication at Discharge
0117 : Beta Blockade at Discharge
0118 : Anti-Lipid Treatment Discharge
0119 : Risk-Adjusted Operative Mortality for CABG
0120 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)
0121 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement

0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery
0127 : Preoperative Beta Blockade
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
0134 : Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)
1501 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery
2514 : Risk-Adjusted Coronary Artery Bypass Graft (CABG) Readmission Rate

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**
N/A

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?
Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**
N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** STS_Surgeon_Composite_Appendix_-_S.4-S.11-S.14-S.15-_1b.2-_1b.4-_manuscripts.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons

Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
- Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 **Year the measure was first released:** 2015

Ad.3 **Month and Year of most recent revision:** 06, 2016

Ad.4 **What is your frequency for review/update of this measure?** Annually

Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A

Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 3031
Measure Title: STS Mitral Valve Repair/Replacement (MVRR) Composite Score
Measure Steward: The Society of Thoracic Surgeons
Brief Description of Measure: The STS Mitral Valve Repair/Replacement (MVRR) Composite Score measures surgical performance for isolated MVRR with or without concomitant tricuspid valve repair (TVr), surgical ablation for atrial fibrillation (AF), or repair of atrial septal defect (ASD). To assess overall quality, the STS MVRR Composite Score comprises two domains consisting of six measures:
Domain 1 – Absence of Operative Mortality
Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.
Domain 2 – Absence of Major Morbidity
Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Outcome data are collected on all patients and from all participants. For optimal measure reliability, participants meeting a volume threshold of at least 36 cases over 3 years (i.e., approximately one mitral case per month) receive a score for each of the two domains, plus an overall composite score. The overall composite score is created by "rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by the following:
1 star – lower-than-expected performance
2 stars – as-expected performance
3 stars – higher-than-expected performance

Developer Rationale: Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, mortality alone fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events. This composite provides a more comprehensive measure of overall quality and is timely due to the fact that mitral valve operations are being performed with increasing frequency for a variety of etiologies and pathologies.

Numerator Statement: See appendix

Denominator Statement: See appendix
Denominator Exclusions: See appendix

Measure Type: Composite
Data Source: Electronic Clinical Data: Registry
Level of Analysis: Clinical: Group/Practice, Facility

New Measure -- Preliminary Analysis

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

- This new composite measure of healthcare outcomes is comprised of an absence of operative mortality domain and an absence of major morbidity domain that includes any one or more of the identified complications.
- It is based on 7 NQF-endorsed measures of which 2 are mortality measures and 5 are cardiac surgery-related major morbidities. The developer reports that mortality alone does not take into account that not all operative survivors receive equal care and notes that mitral valve operations are being performed with increasing frequency. The 7 measures are:
  - 0121: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement
  - 1501: Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
  - 0114: Risk-Adjusted Postoperative Renal Failure
  - 0115: Risk-Adjusted Surgical Re-exploration
  - 0129: Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
  - 0130: Risk-Adjusted Deep Sternal Wound Infection
  - 0131: Risk-Adjusted Stroke/Cerebrovascular Accident

- NQF criteria indicate that each component in a composite must meet the evidence sub criterion to justify its inclusion in the composite and be NQF-endorsed or evaluated as meeting measure evaluation criteria.
- The components of this composite are outcomes for which the required evidence is identification of a relationship between the outcome and at least one healthcare action that could achieve change in measure results. Information regarding service and/or care to impact mortality and 4 of the 5 morbidities is provided.
- The developer states that by taking into account major morbidity, it provides a more comprehensive measure of overall quality. It also notes that mitral valve operations are being performed with increasing frequency.
- References that address operative morbidity and mortality dating from 1998 through 2016 are provided.
- Approach to the work (see appendix S.14 and S.15) that underpins the measure is described and references provided. This includes information about modification of existing mitral valve models to adjust for case mix (consistent with evolving trends).
- References that address operative mortality and morbidity dating from the 1990's through 2014, including those related to current STS adult cardiac surgery risk models, are provided.

Question for the Committee:

- *Is the information regarding modifications and application of the model to the development of the composite clear and compelling?*
- *Does the Committee agree that the components together convey an appropriate measure of mitral valve surgery quality?*

- *Is there at least one thing that the provider can do to achieve a change in measure results?*

Guidance from the Evidence Algorithm: Assess performance on outcome (Box 1) – Relationship between outcome and healthcare action (Box 2)

Preliminary rating for evidence: ☒ Pass ☐ No Pass

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

- For this new measure, calculation was done using STS data for patients undergoing isolated mitral valve repair/replacement in 2 consecutive overlapping 3 year periods (July 2011 – June 2014 and July 2012 – June 2015).
  - The developer reports that 3.5% and 3.4% of STS participants with ≥ 36 cases in the respective time periods have lower than expected performance on the measure based on 95% Bayesian credible interval. In comparison, 2.7% and 2.2% of all participants have lower than expected performance.

Disparities (see Appendix 1b4)

Logistic regression was used to study associations of race (black), ethnicity (Hispanic), and insurance status (among patients < and ≥ age 65) with operative mortality and major morbidity while adjusting for the measure's risk adjustment model covariates. Odds ratios with 95% confidence intervals and p-values are summarized.

	OR (95% CI)	**P**	**OR (95% CI)**	**P**
Insurance status among patients age ≥65				
Medicare without Medicaid/Commercial	(ref)		(ref)	
Medicare and Medicaid dual eligible	1.05 (0.90, 1.21)	0.5415	0.93 (0.71, 1.22)	0.0806
Medicare Commercial without Medicaid	0.96 (0.87, 1.05)	0.3875	1.06 (0.91, 1.24)	0.2377
Insurance Status among patients age <65				
Commercial or HMO without Medicare/Medicaid	(ref)		(ref)	
Medicare or Medicaid	1.26 (1.16, 1.38)	<.0001	1.43 (1.16, 1.76)	0.0007
None/Self Pay	1.21 (1.06, 1.39)	0.0047	1.20 (0.86, 1.68)	0.2912
Other	1.23 (1.03, 1.47)	0.0223	1.68 (1.11, 2.53)	0.0135
Black race	1.21 (1.12, 1.31)	<.0001	0.93 (0.78, 1.10)	0.3999
Hispanic Ethnicity	1.20 (1.07, 1.35)	0.0023	1.16 (0.93, 1.45)	0.1877

Questions for the Committee:

- *In considering whether there is a gap in care that warrants a national performance measure, does the fact that each component of the measure represents occurrence of a serious adverse (never) event influence Committee thinking?*
- *Is handling of disparities data clear and are you aware of evidence that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

1c. Composite - Quality Construct and Rationale
Maintenance measures – same emphasis on quality construct and rationale as for new measures.

1c. Composite Quality Construct and Rationale. The quality construct and rationale should be explicitly articulated and logical; a description of how the aggregation and weighting of the components is consistent with the quality construct and rationale also should be explicitly articulated and logical.

- The development of the measure, the model upon which it is based and the modifications to the model (see Badhwar et al) to adjust for case mix are discussed in detail in the measure submission.
- The measure is based on a combination of NQF-endorsed risk-adjusted operative mortality outcome measures and 5

risk-adjusted major complications. The developer states that an NQF-endorsed structure measure, database participation (0113), is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive scores. It differs from the NQF-endorsed STS CABG Composite Score in that it does not include process measures due to the fact that for MVRR, no widely accepted process measures that meet performance metric criteria currently exist.

- The composite comprises 2 domains.
  - Domain 1 includes the proportion of patients (risk-adjusted) who do not experience operative mortality (death before hospital discharge or within 30 days of operation).
  - Domain 2 includes proportion of patients (risk-adjusted) who do not experience any major morbidity (occurrence of any one or more of 1) prolonged ventilation; 2) deep sternal wound infection; 3) permanent stroke; 4) renal failure; and 5) reoperations for bleeding, prosthetic or native valve dysfunction, and other cardiac reasons, but not for non-cardiac reasons).
  - Participants receive a score for each of the 2 domains plus an overall composite score.
- The developer states that average mortality rates for the procedures of interest are at very low levels making differentiating performance based on mortality alone difficult in that it fails to take into account the fact that not all operative survivors received equal quality care.
- The mortality domain corresponds to a single measure. The study endpoint for the morbidity domain combines multiple measures. Mortality rates were converted to survival rates and morbidity rates were converted to "absence of morbidity" rates. The developer notes that defining scores in this manner ensures that increasingly positive values reflect better performance. The overall composite score is created by "rolling up" the domain scores into a single number.
- Aggregation and weighting of the composite components is described. The mortality and morbidity rates were converted to survival and absence of morbidity rates, respectively. The weighting was assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the 2 domains.
- The minimum threshold for receiving a site-specific STS MVRR composite score (see Badhwar et al) (1 mitral case per month or 36 cases over 3 years) was selected on the basis of expert opinion and statistical testing – reliability = 0.58 (95% probability interval).

Questions for the Committee:

- *Are the quality construct and a rationale for the composite explicitly stated and logical?*
- *Is the method for aggregation and weighting of the components explicitly stated and logical?*
- *Does the Committee agree that the adaptation of NQF-endorsed measures meets the expectation that individual components of a composite meet NQF criteria?*

Preliminary rating for composite quality construct and rationale:

☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

Note: Qualifies for high rating if Committee agrees that NQF expectation regarding endorsement or evaluation of component measures is satisfied.

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- new composite outcome (periop complications/death) measure-logical association w quality
- same as with CABG.
- participants with fewer cases appear to do better ?
- circa 10% low and high outliers"
- This measurement is a composite of clinical outcome measure after a well-defined cardiac surgical procedure. This procedure is increasingly being performed and most frequently in Medicare aged population. Many of the outcome end points are high risk complications for the patient and represent significant resource utilization if they occur. Measurement would be useful to inform patients and policymakers

1b.

- Performance among database participants with at least 36 cases over 3 years during 2 time periods (July 2011 – June 2014 and July 2012 – June 2015 is provided:

- 397 (85.9%) and 400 (85.1%) performed as-expected during the 2 respective time periods
- 16 (3.5%) and 16 (3.4%) had lower-than-expected performance;
- 49 (10.6%) and 54 (11.5%) had higher-than-expected performance.
- Also, performance data for all participants, including those with less than 36 cases, is provided.
- Using 2012 – 2015 data, risk-adjusted mortality and morbidity rates across the 3 performance categories based on 2011 – 2014 data were compared. Results were:
- 3.2% mortality and 16.9% morbidity among participants with as-expected performance;
- 6.1% mortality and 27.4% morbidity among participants with lower-than-expected performance;
- 1.5% mortality and 11.1% morbidity among participants with higher-than-expected performance.
- Developer interpretation of the results is that the composite measure behaves as expected and results are reasonably consistent across 2 consecutive overlapping time periods
- By using the STS database there is noted variation in outcomes in those facilities that meet the required volume threshold of 36 cases/3 years. The developers should provide more information on from their dataset on exactly how many facilities meet this measure and the number that do not.

1c.

- well constructed
- This is a composite measure that including mortality and major morbidities. These outcomes have been well defined and previously endorsed by the NQF committee as they are part of other STS-cardiac measures. Their inclusion is appropriate as they represent the major complications associated with this procedure.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

- The data source for the measure is the STS Adult Cardiac Surgery Database. Data collection occurs through an electronic system using a detailed collection tool.
- The measure is specified for analysis at the group/practice level and is intended for use at the hospital/acute care setting.
- The composite score provides surgical performance for isolated MVRR with or without concomitant tricuspid valve repair, surgical ablation for atrial fibrillation, or repair of atrial septal defect. It is comprised of two domains: absence of operative mortality and absence of major morbidity (described under Construct). The measure is based on a combination of risk-adjusted operative mortality for isolated mitral valve repair or replacement (MVRR) and risk-adjusted occurrence of any one or more of 5 major complications.
- To adjust for case mix in the MVRR Composite Score (see Badhwar et al in appendix), the published 2008 STS isolated valve model (see O'brien in appendix) was modified and re-estimated in the current study population. The developer notes that the main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population that did not meet inclusion/exclusion criteria for the 2008 model.
- Details regarding development of the risk-adjustment model and the approach to scoring are described in the measure submission and in an "article in press" by Badhwar et al.

Questions for the Committee :

- *Is there any question regarding whether the measure can be consistently abstracted from electronic or paper records by non-STS registry members?*

2a2. Reliability Testing Testing attachment

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is

precise enough to distinguish differences in performance across providers.

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

- The measure was developed and tested using data from the STS Adult Cardiac Surgery Database for patients from 867 database participants undergoing MVRR during July 2011 – June 2014.
- Risk model discrimination and calibration was assessed using data from 62,118 eligible patients undergoing MVRR during July 2011 – June 2014 from 1,064 database participants.
- The developer notes that to ensure adequate statistical precision, composite scores only for participants with at least 36 eligible cases during the 3 year measurement window will be reported.
- Estimated reliability of the measure using 3 years of data from participants with at least 36 total cases was 0.58.
- The mathematical approach to signal-to-noise estimation is detailed.

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the rationale for selection of the threshold for reporting clear and acceptable?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm: Precise specifications (Box 1) – Empirical testing (Box 2) – Testing with measure score (Box 4) – Testing method described (Box 5) – Confidence that scores are reliable (Box 6)

Preliminary rating for reliability: ☒ **High** ☐ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☒ **Yes** ☐ **Somewhat** ☐ **No**

Question for the Committee:

- *Does the Committee agree that the specifications are consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Describe any updates to validity testing: N/A

SUMMARY OF TESTING

Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:

☒ **Face validity only**

☐ **Empirical validity testing of the measure score**

Validity testing method:

- Using the concept of performance categories, risk-adjusted mortality and morbidity rates were compared across 3 performance groups.
- The extent to which a participant's composite score remained stable across 2 consecutive reporting periods was assessed.

- Degree of uncertainty around a database participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CIs).
- Point estimates and CIs for an individual participant are reported with a comparison to benchmarks (overall average STS composite score and several percentiles) based on the national sample.
- Also, the composite measure result is converted into groups or categories using a Bayesian CI. If the CI around the composite score overlaps the overall STS average, the participant is performing at the as-expected level (indistinguishable from the average). If the CI is entirely above the national average, the participant has "higher-than-expected" performance and if entirely below the STS national average, the participants has "lower-than-expected" performance.

Validity testing results:

- Performance among database participants with at least 36 cases over 3 years during 2 time periods (July 2011 – June 2014 and July 2012 – June 2015 is provided:
  - 397 (85.9%) and 400 (85.1%) performed as-expected during the 2 respective time periods;
  - 16 (3.5%) and 16 (3.4%) had lower-than-expected performance;
  - 49 (10.6%) and 54 (11.5%) had higher-than-expected performance.
  - Also, performance data for all participants, including those with less than 36 cases, is provided.
- Using 2012 – 2015 data, risk-adjusted mortality and morbidity rates across the 3 performance categories based on 2011 – 2014 data were compared. Results were:
  - 3.2% mortality and 16.9% morbidity among participants with as-expected performance;
  - 6.1% mortality and 27.4% morbidity among participants with lower-than-expected performance;
  - 1.5% mortality and 11.1% morbidity among participants with higher-than-expected performance.
- Developer interpretation of the results is that the composite measure behaves as expected and results are reasonably consistent across 2 consecutive overlapping time periods

Questions for the Committee:

- *Is the work to demonstrate validity clear and complete such that you can agree that the score from this measure as specified is an indicator of quality?*
- *Do the results demonstrate sufficient validity so that conclusions about quality can be made?*

2b3-2b7. Threats to Validity

2b3. Exclusions:

There are no exclusions.

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☒ **Yes** ☐ **No**

- Associations of race, ethnicity and insurance (see appendix 1b.4) status with operative mortality and major morbidity were studied using logistic regression and results provided. Race (black) and ethnicity (Hispanic) are included in the STS 2008 models.
- As noted in the reliability section, the published STS 2008 isolated valve model was modified and re-estimated in the current study population, primarily to be able to calculate risk estimates for patients in the current study population who did not meet inclusion/exclusion criteria for the existing isolated valve model.
  - Covariates for the modified mortality model were identical to the 2008 model and covariates were identical to the mortality or major morbidity model except:
    - Adjustment variable for concomitant tricuspid repair was not in the 2008 model so was included;

    - Adjustment for tricuspid insufficiency using redefined categories of none or mild, moderate, and severe;
    - Adjustment for infectious endocarditis to provide for a treated category not included in the 2008 model.
  - Estimated odds ratios from the modified STS 2008 models are provided.
- Discrimination and calibration was assessed using data from 62,118 patients undergoing MVRR during July 2011 – June 2014.
  - Discrimination was gauged by calculating C-statistics for both models. Bootstrapping was used to estimate and adjust for "optimism" from estimating and evaluating the model on the same sample. Bootstrap-adjusted estimated C-statistic was 0.746 for morbidity model and 0.807 for mortality model.
  - Calibration was evaluated using 5-fold cross validation. The approach is described in some detail. Expected to observed plots across the 5 samples are provided.
  - The stated conclusion is that the risk models are well calibrated and have good discrimination power.

Questions for the Committee:

- *Does the approach and outcomes of modification of the STS 2008 model demonstrate an appropriate risk-adjustment strategy for the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Are there SDS factors that should be considered for evaluation as the measure is implemented?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

- As noted in the validity testing section, performance among database participants with at least 36 cases over 3 years during 2 time periods (July 2011 – June 2014 and July 2012 – June 2015) demonstrate differences among 3 groupings:
  - 397 (85.9%) and 400 (85.1%) performed as-expected during the 2 time periods;
  - 16 (3.5%) and 16 (3.4%) had lower-than-expected performance;
  - 49 (10.6%) and 54 (11.5%) had higher-than-expected performance.

Question for the Committee:

- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Not needed. Single data source.

2b7. Missing Data

- The overall frequency of missing data is reported at 0.5% for operative mortality and 0.4% for major complications.
  - Median participant-specific frequency of missing data was 0% (range 0% to 69%) for mortality and 0% (range 0% to 38%) for major complications.
  - Percent of participants with >10% missing data was 1.0% for mortality and 1.3% for major complications.
  - Participant-specific mortality and complication rates, after excluding records with missing data from the denominator, were calculated as a sensitivity analysis. There was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.

2d. Composite measure: construction

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

- Correlation between each domain-specific estimate and overall composite score was calculated. Pearson correlations were 0.74 for mortality versus overall composite measure and 0.89 for morbidity domain score versus overall score.
- To form the composite, morbidity and mortality domains were rescaled by dividing by their respective standard deviations across STS participants and then adding the two domains together. The weighting was assessed by an expert panel to determine if an appropriate reflection of the relative importance of the 2 domains was provided. Relative weights in the final composite of risk-standardized mortality and risk standardized major morbidity were 0.74 and 0.26 respectively. This weighting was consistent with the expert panel's clinical assessment of each domain's relative importance.

Questions for the Committee:

- *Do the component measures fit the quality construct?*
- *Are the objectives of parsimony and simplicity achieved while supporting the quality construct?*

Guidance from the Validity Algorithm : Measure specification consistent with evidence (Box 1) – Potential threats to validity (Box 2) – Empirical validity testing (Box 3) – Face validity testing (Box 4) – Agreement that score can be used to distinguish quality (Box 5) – Moderate

Preliminary rating for validity: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.

- clear construct & implementable
- No concerns. Risk-adjustment is based upon the detailed analysis of the STs database and expert consensus.

2a2.

- reliable
- Reliability is based upon the volume threshold of 36 cases/3 years. The reliability value is reported as 0.58 which is moderate reliability.

2b1.

- consistent & valid
- No concerns

2b2.

- valid
- The committee has previously endorsed these elements as being valid clinically important outcomes appropriate for wide-spread use in measurement.

2d.

- yes
- The composite construct of weighting mortality versus the major morbidity was performed by an expert panel which heavily weighted mortality (0.740) versus morbidity (0.26). Previously, the developers noted that mortality is a very low risk and the major complications represent the significant resource drain. I would ask the developers to provide more information on how this weighting was achieved and how the model responds if the relative weightings are changed.

Criterion 3. Feasibility

Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

- The developer states that required data elements are generated or collected and used by healthcare personnel during provision of care. They are then abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction). Some of the elements are available in EHRs or from other electronic sources.
- Per the developer, the data elements in the measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some for more than 20 years. The database has more than 1,100 participants (representing over 90% of programs that provide cardiac surgery in US). Local availability of data elements will vary from full EHR capability to no availability; however, all data elements are submitted to the STS database in electronic format following a standard set of data specifications.
- There are no additional costs for data collection specific to the measure. Costs to develop and maintain the measure include volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.
- STS Adult Cardiac Surgery Database participants (single or group of surgeons) pay annual participant fees of $3,500 if majority of surgeons in the group are STS members and $4,750 if the majority are not STS members. In addition there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

Questions for the Committee:
Is the effort and cost associated with abstracting the required data elements appropriate to the value of the measure?

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

- Feasible
- No concerns.

Criterion 4: Usability and Use

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☐ **Yes** ☒ **No**

Current use in an accountability program? ☐ **Yes** ☐ **No**
OR
Planned use in an accountability program? ☒ **Yes** ☐ **No**

This new composite measure was developed in 2015 and will be published in 2016. STS plans to distribute participant-specific results in 2016 and begin public reporting "within the next year or so".

Questions for the Committee*:*
- *Does the Committee have concern about any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Public reporting planned "year or two"
- This could be very useful but it really depends on the number of facilities that meet this threshold case volume. Using the test set data that was presented to demonstrate performance variation there are somewhere around 800 or so facilities but what is not clear is how many facilities do these cases below the threshold?? It may not be meaningful if it only impact 20-30% of institutions.

Criterion 5: Related and Competing Measures

Related or competing measures

Related measures include STS measures that have been included in development of the composite or are otherwise related. They are harmonized.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** Click here to enter NQF number
Measure Title: STS Mitral Valve Repair/Replacement (MVRR) Composite Score
IF the measure is a component in a composite performance measure, provide the title of the Composite Measure here: Click here to enter composite measure #/ title

Date of Submission: 6/5/2016

Instructions

- *For composite performance measures:*
  - *A separate evidence form is required for each component measure unless several components were studied together.*
  - *If a component measure is submitted as an individual performance measure, attach the evidence form to the individual measure submission.*
- Respond to all questions as instructed with answers immediately following the question. All information needed to demonstrate meeting the evidence subcriterion (1a) must be in this form. An appendix of *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 10 pages (*incudes questions/instructions*; minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the evidence for this measure meets NQF's evaluation criteria.

1a. Evidence to Support the Measure Focus

The measure focus is evidence-based, demonstrated as follows:

- Health outcome: [3] a rationale supports the relationship of the health outcome to processes or structures of care. Applies to patient-reported outcomes (PRO), including health-related quality of life/functional status, symptom/symptom burden, experience with care, health-related

behavior.

- Intermediate clinical outcome: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured intermediate clinical outcome leads to a desired health outcome.
- Process: [5] a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured process leads to a desired health outcome.
- Structure: a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence [4] that the measured structure leads to a desired health outcome.
- Efficiency: [6] evidence not required for the resource use component.

Notes

3. Generally, rare event outcomes do not provide adequate information for improvement or discrimination; however, serious reportable events that are compared to zero are appropriate outcomes for public reporting and quality improvement.

4. The preferred systems for grading the evidence are the U.S. Preventive Services Task Force (USPSTF) grading definitions and methods, or Grading of Recommendations, Assessment, Development and Evaluation (GRADE) guidelines.

5. Clinical care processes typically include multiple steps: assess → identify problem/potential problem → choose/plan intervention (with patient input) → provide intervention → evaluate impact on health status. If the measure focus is one step in such a multistep process, the step with the strongest evidence for the link to the desired outcome should be selected as the focus of measurement. Note: A measure focused only on collecting PROM data is not a PRO-PM.

6. Measures of efficiency combine the concepts of resource use and quality (see NQF's Measurement Framework: Evaluating Efficiency Across Episodes of Care; AQA Principles of Efficiency Measures).

1a.1.This is a measure of: (*should be consistent with type of measure entered in De.1*)

Outcome

☒ Health outcome: 1. Operative Mortality; 2. Postoperative Major Morbidity

☐ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors

☐ Intermediate clinical outcome (*e.g., lab value*): Click here to name the intermediate outcome

☐ Process:

☐ Structure: Click here to name the structure

☐ Other: Click here to name what is being measured

_____

HEALTH OUTCOME/PRO PERFORMANCE MEASURE *If not a health outcome or PRO, skip to 1a.3*

1a.2. Briefly state or diagram the path between the health outcome (or PRO) and the healthcare structures, processes, interventions, or services that influence it.

Operative Mortality

Mortality likely is the single most important negative outcome associated with a surgical procedure. Operative mortality, defined as death before hospital discharge or within 30 days of the operation, should include nearly all deaths that occur as a direct result of the surgery or an immediate postoperative complication. Critical evaluation of operative mortality allows one to evaluate the risk associated with a given procedure for various patient characteristics, and more importantly, aggressively search for ways to minimize that risk.

Major Morbidity

- Surgical re-exploration for bleeding remains a known complication following cardiac surgery. The literature documents that bleeding following coronary artery bypass surgery confers greater ICU stay and therefore greater resource consumption. It remains unknown and controversial whether long-term outcomes are worse for the isolated re-exploration for bleeding patients. However, Hein documents that patients with ICU stay > 3 days (with bleeding as multivariate risk factor for this outcome), have a long-term survival which is inferior to patients with ICU stay < 3 days. The patient consequences of this complication relates to the physiological stress of facing another operation and receiving blood products.

- A devastating complication of cardiac surgery is deep sternal wound infection. Patients with deep sternal wound infection require multiple surgeries to clear the infection, have longer hospital stays, greatly increased costs and increased early and late mortality. Care processes that influence the incidence of sternal infection span the first 3 major phases of surgical care. In the preoperative phase, routine patient decontamination and identification of active infections are key elements. In the intraoperative phase, impeccable surgical aseptic technique and timing of antibiotic administration are key processes. In the postoperative phase, aseptic wound care and detection of early wound inflammation are important in preventing delayed contamination and subsequent infection.
- Prolonged ventilation has been shown to substantially increase length of stay, the costs of care, and is associated with higher rates of respiratory failure, stroke, renal failure, and death. Modalities to decrease the rate of prolonged intubation include physician supervised protocols for extubation implemented by nurses and respiratory therapists, improved preoperative preparation of patients, reduction of postoperative bleeding, and intra-operative protocolized anesthesia care. Current implementation is highly variable and great opportunities to increase the implementation of evidence based care exist. Cardiac surgery programs with high implementation have lower than average rates of prolonged ventilation and significantly lower rates of adverse events.
- Postoperative renal failure is an occasional but serious complication in the cardiac surgical population and is a major determinant of short- and long-term survival. Identification of clinical precursors of postoperative renal insufficiency and improvement in perioperative treatment of this high-risk group will improve the long-term survival of our patients. By implementing known recommendations (delay heart surgery, when possible, after cardiac catheterization, maintain mean CPB perfusion pressure at 80% of preop BP, etc), postoperative kidney injury should be significantly reduced.
- Postoperative stroke/CVA produces significant short- and long-term often devastating effects to patients and their families. It is associated with significant increases in death, respiratory failure, renal failure, length of stay, and cost of care. Modalities to decrease the rate of postoperative stroke include measures to maintain blood pressure and perfusion, glycemic control, avoidance of atrial fibrillation, anticoagulation protocols, etc. Many opportunities exist to decrease stroke rates by increasing implementation of evidence based strategies.

References – Operative Mortality

- Birkmeyer NJ, Marrin CA, et al. Decreasing mortality for aortic and mitral valve surgery in Northern New England. Northern New England Cardiovascular Disease Study Group. *Ann Thorac Surg.* 2000;70(2):432-437.
- Edwards FH, Petyerson ED, et al. Prediction of operative mortality following valve replacement surgery. *JACC.* 37:3:885-892.
- Goodney PP, O'Connor GT, et al. Do hospitals with low mortality rates in coronary artery bypass also perform well in valve replacement? *Ann Thorac Surg.* 2003;76:1131-1137.
- Mehta RH, Eagle KA, et al. Influence of age on outcomes in patients undergoing mitral valve replacement. *Ann Thorac Surg.* 2002;74:1459-1467.
- Iribarne A, Russo MJ, Easterwood R et al. Minimally invasive versus sternotomy approach for mitral valve surgery: a propensity analysis. *Ann Thorac Surg.* 2010;90:1471–1477
- LaPar DJ, Hennessy S, Fonner E, et al. Does urgent or emergent status influence choice in mitral valve operations? An analysis of outcomes from the Virginia Cardiac Surgery Quality Initiative. 2010;90:153-60
- Umakanthan R, Petracek MR, Leacche M et al, Minimally invasive right lateral thoracotomy without aortic cross-clamping: an attractive alternative to repeat sternotomy for reoperative mitral valve surger;y. J Heart Valve Dis. 2010;19:236-43
- Vassileva CM, McNeely C, Spertus J, Markwell S, Hazelrigg S. Hospital volume, mitral repair rates, and mortality in mitral valve surgery in the elderly: An analysis of US hospitals treating Medicare fee-for-service patients. J Thorac Cardiovasc Surg. 2014pii: S0022-5223(14)01290-2
- Chatterjee S, Rankin JS, Gammie JS, et al. Isolated mitral valve surgery risk in 77,836 patients from the Society of Thoracic Surgeons database. Ann Thorac Surg. 2013;96:1587-94
- LaPar DJ, Ailawadi G, Isbell JM, et al. Virginia Cardiac Surgery Quality Initiative. Mitral valve repair rates correlate with surgeon and institutional experience. J Thorac Cardiovasc Surg. 2014;148:995-1003

- Dayan V, Soca G, et al. Similar survival after mitral valve replacement or repair for ischemic mitral regurgitation: a meta-analysis. Ann Thorac Surg. 2014 Mar;97(3):758-65.
- Kaneko T, Aranki S, et al. Mechanical versus bioprosthetic mitral valve replacement in patients <65 years old. J Thorac Cardiovasc Surg. 2014 Jan;147(1):117-26.

References – Major Morbidity

- Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *Ann Thorac Surg.* 2003;75:1856-1865.
- Hein OV, Birnbaum J, Wernecke K, England M, Knoertz W, Spies C. Prolonged Intensive Care Unit Stay in Cardiac Surgery: Risk Factors and Long-Term Survival. *Ann Thor Surg* 2006;81:880-85.
- Stamou SC, Camp SL, Stiegel RM, et al. Quality improvement program decreases mortality after cardiac surgery. *J Thorac Cardiovasc Surg* 2008;136:494-499.
- Braxton JH, Marrin CA, McGrath PD, et al. 10-Year follow-up of patients with and without mediastinitis. Semin Thorac Cardiovasc Surg. 2004;16:70–76.
- Graf K, Ott E, Vonberg RP, et al. Economic aspects of deep sternal wound infections. Eur J Cardiothorac Surg 2010;37:893-96.
- Edwards FH, Engelman RM, Houck P et al. The Society of Thoracic Surgeons Practice Guideline Series: Antibiotic Prophylaxis in Cardiac Surgery, Part I: Duration. Ann Thorac Surg 2006;81: 397-404,
- Wilson APL, Gibbons C, Reeves BC, et al. Surgical wound infection as a performance indicator: agreement of common definitions of wound infection in 4773 patients. BMJ 2004;329:720-24.
- Filsoufi F, Castillo JG, Rahmanian PB, et al. Epidemiology of deep sternal wound infection in cardiac surgery. J Cardiothorac Vasc Anesth 2009;23:488-94.
- Koch CG, Nowicki ER, Rajeswaran J, et al. When the timing is right: antibiotic timing and infection after cardiac surgery. J Thorac Cardiovasc Surg 2012;144:931-37.
- Paul M, Raz, A, Leibovici L, et al. Sternal wound infection after coronary artery bypass graft surgery: validation of existing risk scores. J Thorac Cardiovasc Surg 2007;133:397-403.
- Lazar HL, Ketchedjian A, Haime M, et al. Topical Vancomycin in combination with perioperative antibiotics and tight glycemic control helps to eliminate sternal wound infections. J Thorac Cardiovasc Surg 2014;148:1035-40.
- Miyahara K, MatsuuraA, Takemura H, et al. Implementation of bundled interventions greatly decreases deep sternal wound infection following cardiovascular surgery. J Thorac Cardiovasc Surg 2014;148:2381-88.
- Matros E, Aranki, SF, Bayer LR, et al. Reduction in incidence of deep sternal wound infections: random or real? J Thorac Cardiovasc Surg 2010;139:680-85.
- Bardell T, Legare JF, Buth KJ, et al. ICU readmission after cardiac surgery. Eur J Cardiothorac Surg. 2003;23(3):354-359.
- Meade MO, Guyatt G, Butler R, et al. Trials comparing early vs late extubation following cardiovascular surgery. Chest. 2001:120(6 Suppl):445S-453S.
- Naughton C, Reilly N, Powroznyk A, et al. Factors determining the duration of tracheal intubation in cardiac surgery: a single-centre sequential patient audit. Eur J Anaesthesiol. 2003;20(3):225-233.
- Welke KF, Ferguson TB, Coombs LP, et al. Validity of the Society of Thoracic Surgeons National Adult Cardiac Surgery Database. Ann Thorac Surg. 2004;77:1137-1139.
- Knapik P, Ciesla D, Borowik D, Czempik P, Knapik T. Prolonged ventilation post cardiac surgery – tips and pitfalls of the prediction game. J Cardiothorac Surg 2011;6:158.
- Alsabbagh MM, Asmar A, Ejaz NI, Aiyer RK, Kambhampati G, Ejaz AA. Update on clinical trials for the prevention of acute kidney injury in patients undergoing cardiac surgery. Am J Surg 2013;206:86-95
- Arora P, Kolli, H, Nainani N, Nader N, Lohr J. Preventable risk factors for acute kidney injury in patients undergoing cardiac surgery. J Cardiothorac Vasc Anesth 2012; 26:687-697.
- Boldt J, Brenner T, Lehmann A, Suttner SW, Kumle B, Isgro F. Is kidney function altered by the duration of cardiopulmonary bypass? Ann Thorac Surg. 2003;75(3):906-912.
- Chertow GM, Levy EM, Hammermeister KE, et al. Independent association between acute renal failure and mortality following cardiac surgery. Am J Med. 1998;104(4):343-348

- Conlon PJ, Stafford-Smith M, White WD, Newman MF, King S, Winn MP, Landolfo K. Acute renal failure following cardiac surgery. Nephrol Dial Transplant. 1999;14(5):1158-1162.
- Haase M, Haase-Fielitz A, Bellomo R, Devarajan P, Story D, Matalanis G, Reade MC, Bagshaw SM, Seevanayagam N, Seevanayagam S, Doolan L, Buxton B, Dragun D. Sodium bicarbonate to prevent increases in serum creatinine after cardiac surgery: a pilot double-blind, randomized trial. Crit Care Ned 2009;37:39-47.
- Kramer RS, Quinn RD, Groom RC, Braxton JH, Malenka DJ, Kellett MA, Brown JR for the Northern New England Cardiovascular Disease Study Group. Same admission cardiac catheterization and cardiac surgery: is there an increased incidence of acute kidney injury? Ann Thorac Surg 2010;90:1418-1424.
- Mangano CM, Diamondstone LS, Ramsay JG, et al. Renal dysfunction after myocardial revascularization: risk factors, adverse outcomes, and hospital resource utilization: the Multicenter Study of Perioperative Ischemia Research Group. Ann Intern Med. 1998;128(3):194-203.
- Ranucci M, Ballotta A, Agnelli B, Frigiola A, Mencanti L, Castelvecchio S, for the Surgical and Clinical Outcome Research (SCORE) Group. Acute kidney injury in patients undergoing cardiac surgery and coronary angiography on the same day. Ann Thorac Surg 2103;95:513-519.
- Rosner MH, Okusa MD. Acute kidney injury associated with cardiac surgery. Clin J Am Soc Nephrol 2006;1:19-32.
- Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SLT, DeLong ER, O'Brien SM, Shewan CM, Dokholyan RS, Peterson ED. Quality Measurement in adult cardiac surgery: Part 1-conceptual framework and measure selection. Ann Thorac Surg 2007;83:S3-S12
- Tang AT, Alexiou C, Hsu J, Sheppard SV, Haw MP, Ohri SK. Leukodepletion reduces renal injury in coronary revascularization: a prospective randomized study. Ann Thorac Surg. 2002;74(2):372-327; discussion 377.
- Amory DW, Grigore A, Amory JK, et al. Neuroprotection is associated with beta-adrenergic receptor antagonists during cardiac surgery: evidence from 1,575 patients. J Cardiothorac Vasc Anesth. 2002;16(3):270-277.
- Arsenault KA, Yusus AM, Crystal E, Healey JS, Morillo CA, Nair GM et al. Interventions for preventing postoperative atrial fibrillation in patients undergoing heart surgery. Cocrane Database Syst Rev. 2013; 1:CD003611
- Bucerius J, Gummert JF, Borger MA, et al. Predictors of delirium after cardiac surgery delirium: effect of beating-heart (off-pump) surgery. J Thorac Cardiovasc Surg. 2004;127(1):57-64.
- Inoue K, Luth JU, Pottkamper D, et al. Incidence and risk factors of perioperative cerebral complications: heart transplantation compared to coronary artery bypass grafting and valve surgery. J Cardiovasc Surg. 1998;39(2):201-208.
- Rosenberger P, Shernan SK, Loffler M, Shekar PS, Fox JA, Tuli JK, Nowak M and Eltzschig HK. The influence of epiaortic ultrasonograpy n intraoperative surgical management in 6051 cardiac surgical patients. Ann Thorac Surg. 2008; 85: 548-53.

1a.2.1. State the rationale supporting the relationship between the health outcome (or PRO) to at least one healthcare structure, process, intervention, or service (*i.e., influence on outcome/PRO*).

Please see response above.

Note: For health outcome/PRO performance measures, no further information is required; however, you may provide evidence for any of the structures, processes, interventions, or service identified above.

INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURE

1a.3. Briefly state or diagram the path between structure, process, intermediate outcome, and health outcomes. Include all the steps between the measure focus and the health outcome.

1a.3.1. What is the source of the systematic review of the body of evidence that supports the performance measure?

☐ Clinical Practice Guideline recommendation – ***complete sections 1a.4, and 1a.7***

☐ US Preventive Services Task Force Recommendation – ***complete sections 1a.5 and 1a.7***

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*) – ***complete sections 1a.6 and 1a.7***

☐ Other – ***complete section 1a.8***

Please complete the sections indicated above for the source of evidence. You may skip the sections that do not apply.

_____

1a.4. CLINICAL PRACTICE GUIDELINE RECOMMENDATION

1a.4.1. Guideline citation (*including date*) and **URL for guideline** (*if available online*):

1a.4.2. Identify guideline recommendation number and/or page number and **quote verbatim, the specific guideline recommendation**.

1a.4.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.4.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: If separate grades for the strength of the evidence, report them in section 1a.7.*)

1a.4.5. Citation and URL for methodology for grading recommendations (*if different from 1a.4.1*)**:**

1a.4.6. If guideline is evidence-based (rather than expert opinion), are the details of the quantity, quality, and consistency of the body of evidence available (e.g., evidence tables)?

☐ Yes → ***complete section 1a.7***

☐ No → ***report on another systematic review of the evidence in sections 1a.6 and 1a.7; if another review does not exist, provide what is known from the guideline review of evidence in 1a.7***

_____

1a.5. UNITED STATES PREVENTIVE SERVICES TASK FORCE RECOMMENDATION

1a.5.1. Recommendation citation (*including date*) and **URL for recommendation** (*if available online*):

1a.5.2. Identify recommendation number and/or page number and **quote verbatim, the specific recommendation**.

1a.5.3. Grade assigned to the quoted recommendation with definition of the grade:

1a.5.4. Provide all other grades and associated definitions for recommendations in the grading system. (*Note: the grading system for the evidence should be reported in section 1a.7.*)

1a.5.5. Citation and URL for methodology for grading recommendations (*if different from 1a.5.1*)**:**

Complete section 1a.7

_____

1a.6. OTHER SYSTEMATIC REVIEW OF THE BODY OF EVIDENCE

1a.6.1. Citation (*including date*) and **URL** (*if available online*):

1a.6.2. Citation and URL for methodology for evidence review and grading (*if different from 1a.6.1*)**:**

Complete section 1a.7

1a.7. FINDINGS FROM SYSTEMATIC REVIEW OF BODY OF THE EVIDENCE SUPPORTING THE MEASURE

If more than one systematic review of the evidence is identified above, you may choose to summarize the one (or more) for which the best information is available to provide a summary of the quantity, quality, and consistency of the body of evidence. Be sure to identify which review is the basis of the responses in this section and if more than one, provide a separate response for each review.

1a.7.1. What was the specific structure, treatment, intervention, service, or intermediate outcome addressed in the evidence review?

1a.7.2. Grade assigned for the quality of the quoted evidence with definition of the grade:

1a.7.3. Provide all other grades and associated definitions for strength of the evidence in the grading system.

1a.7.4. What is the time period covered by the body of evidence? (*provide the date range, e.g., 1990-2010*). Date range: Click here to enter date range

QUANTITY AND QUALITY OF BODY OF EVIDENCE

1a.7.5. How many and what type of study designs are included in the body of evidence? (*e.g., 3 randomized controlled trials and 1 observational study*)

1a.7.6. What is the overall quality of evidence across studies in the body of evidence? (*discuss the certainty or confidence in the estimates of effect particularly in relation to study factors such as design flaws, imprecision due to small numbers, indirectness of studies to the measure focus or target population*)

ESTIMATES OF BENEFIT AND CONSISTENCY ACROSS STUDIES IN BODY OF EVIDENCE

1a.7.7. What are the estimates of benefit—magnitude and direction of effect on outcome(s) across studies in the body of evidence? (*e.g., ranges of percentages or odds ratios for improvement/ decline across studies, results of meta-analysis, and statistical significance*)

1a.7.8. What harms were studied and how do they affect the net benefit (benefits over harms)?

UPDATE TO THE SYSTEMATIC REVIEW(S) OF THE BODY OF EVIDENCE

1a.7.9. If new studies have been conducted since the systematic review of the body of evidence, provide for each new study: 1) citation, 2) description, 3) results, 4) impact on conclusions of systematic review.

1a.8 OTHER SOURCE OF EVIDENCE

If source of evidence is NOT from a clinical practice guideline, USPSTF, or systematic review, please describe the evidence on which you are basing the performance measure.

1a.8.1 What process was used to identify the evidence?

1a.8.2. Provide the citation and summary for each piece of evidence.

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
Evidence_Form.STS_MVRR_Composite_Score.docx

1b. Performance Gap

Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:

- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or
- disparities in care across population groups.

1b.1. **Briefly explain the rationale for this measure** *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
N/A

1b.2. **Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.** *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.3. **If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.**
N/A

1b.4. **Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.** *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
See Appendix

1b.5. **If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.**
N/A

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. **Demonstrated high priority aspect of healthcare**
Affects large numbers, A leading cause of morbidity/mortality, Frequently performed procedure, High resource use, Patient/societal consequences of poor quality, Severity of illness

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.
Please see attached evidence form for detailed information.

1c.4. Citations for data demonstrating high priority provided in 1a.3
Please see attached evidence form for list of references.

1c.5. If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful. *(Describe how and from whom their input was obtained.)*
N/A

1d. Composite Quality Construct and Rationale

1d.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.
- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
  - all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient); or
  - any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient).

1d.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1d.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

The STS Mitral Valve Repair/Replacement (MVRR) Composite Score measures surgical performance for isolated MVRR with or without concomitant tricuspid valve repair (TVr), surgical ablation for atrial fibrillation (AF), or repair of atrial septal defect (ASD). Similar to other STS composite measures, this measure is based on a combination of the NQF-endorsed risk-adjusted operative mortality outcome measure and the risk-adjusted occurrence of any of five major complications. An NQF-endorsed structure measure, database participation, is included de facto as only STS Adult Cardiac Surgery Database participants are eligible to receive composite scores. To assess overall quality, the composite comprises the following two domains:

Domain 1 – Absence of Operative Mortality
Proportion of patients (risk-adjusted) who do not experience operative mortality. Operative mortality is defined as death before hospital discharge or within 30 days of the operation.

Domain 2 – Absence of Major Morbidity
Proportion of patients (risk-adjusted) who do not experience any major morbidity. Major morbidity is defined as the occurrence of any one or more of the following major complications:

1. Prolonged ventilation,
2. Deep sternal wound infection,
3. Permanent stroke,
4. Renal failure, and
5. Reoperations for bleeding, prosthetic or native valve dysfunction, and other cardiac reasons, but not for other non-cardiac reasons.

Participants receive a score for each of the two domains, plus an overall composite score. The overall composite score was created by

"rolling up" the domain scores into a single number. In addition to receiving a numeric score, participants are assigned to rating categories designated by one to three stars.

Similar to the NQF-endorsed STS AVR and AVR+CABG measures, the MVRR Composite Score differs from the NQF-endorsed STS CABG Composite Score in that it does not include process measures. This reflects the fact that for MVRR, in comparison with CABG surgery, no widely accepted process measures meeting performance metric criteria currently exist.

1d.3. **Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.**
Risk-adjusted mortality has historically been the dominant outcomes metric for cardiac surgery procedures, but in an era when the average mortality rates for these procedures have declined to very low levels, differentiating performance based on mortality alone is difficult. Specifically, mortality alone fails to take into account the fact that not all operative survivors received equal quality care, e.g., patients who survive surgery but have a debilitating complication that may substantially impact long-term freedom from cardiac events. This composite provides a more comprehensive measure of overall quality and is timely due to the fact that mitral valve operations are being performed with increasing frequency for a variety of etiologies and pathologies.

1d.4. **Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.**
The mortality domain corresponds to a single measure, while the study endpoint for the morbidity domain combines multiple measures and thus is a composite endpoint. To enhance interpretation, mortality rates were converted to survival rates (risk-standardized survival rate = 100 – risk-standardized mortality rate), and morbidity rates were converted to "absence of morbidity" rates (risk-standardized absence of morbidity rate =100 – risk-standardized morbidity rate). Defining scores in this manner ensures that increasingly positive values reflect better performance. The overall composite score is created by "rolling up" the domain scores into a single number.

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Cardiovascular, Surgery, Surgery : Cardiac Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*
Safety, Safety : Complications, Safety : Healthcare Associated Infections

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
http://www.sts.org/sites/default/files/documents/STSAdultCVDataCollectionForm2_73_Annotated.pdf,
http://www.sts.org/sites/default/files/documents/AnnotatedDataCollectionFormV2_81%20April.2015.pdf;

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
N/A

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
See Appendix

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
See Appendix

S.7. **Denominator Statement** *(Brief, narrative description of the target population being measured)*
See Appendix

S.8. **Target Population Category** *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. **Denominator Details** *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See Appendix

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
See Appendix

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
See Appendix

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
N/A

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
See Appendix

S.15. **Detailed risk model specifications** *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at*

measure-specific URL identified in S.1.)
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. **Detailed risk model specifications** (*if not provided in excel or csv file at S.2b*)
See Appendix

S.16. **Type of score:**
Rate/proportion
If other:

S.17. **Interpretation of Score** *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Higher score

S.18. **Calculation Algorithm/Measure Logic** *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
Please see discussion under section S.4 and attached manuscripts.

S.19. **Calculation Algorithm/Measure Logic Diagram URL or Attachment** *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
No diagram provided

S.20. **Sampling** *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.
N/A

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.
N/A

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.5% of records for operative mortality and 0.4% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

The overall frequency of missing data was 0.5% for operative mortality and 0.4% for major complications. The median participant-specific frequency of missing data was 0% (range 0% to 69%) for mortality and 0% (range 0% to 38%) for major complications. The percent of participants with >10% missing data was 1.0% for mortality and 1.3% for major complications. As a sensitivity analysis, we re-calculated participant-specific mortality and complication rates after excluding records with missing data from the denominator. As shown in the figures in section 2b7.2. of the testing attachment, there was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data : Registry

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
STS Adult Cardiac Surgery Database – Version 2.73; STS Adult Cardiac Surgery Database Version 2.81 went live on July 1, 2014.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
Available at measure-specific web page URL identified in S.1

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Clinician : Group/Practice, Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*
Please see section S.4

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
Testing_Form.STS_MVRR_Composite_Score-636008084502899259.docx

NATIONAL QUALITY FORUM—Composite Measure Testing (subcriteria 2a2, 2b2-2b7, 2d)

Measure Number (*if previously endorsed*)**: Click here to enter NQF number**
Composite Measure Title: STS Mitral Valve Repair/Replacement (MVRR) Composite Score
Date of Submission: 6/5/2016
Composite Construction:
☒Two or more individual performance measure scores combined into one score
☐ All-or-none measures (e.g., all essential care processes received or outcomes experienced by each patient)
☒ Any-or-none measures (e.g., any or none of a list of adverse outcomes experienced, or inappropriate or unnecessary care processes received, by each patient)

Instructions: Please contact NQF staff before you begin.

- If a component measure is submitted as an individual performance measure, the non-composite measure testing form must also be completed and attached to the individual measure submission.
- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all composite measures, sections 1, 2a2, 2b2, 2b3, 2b5, and 2d must be completed.**

- **For composites with <u>outcome and resource use</u> measures**, section **2b4** also must be completed.
- If specified for **<u>multiple data sources/sets of specificaitions</u>** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to <u>all</u> questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2), validity (2b2-2b6), and composites (2d) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 25 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

<u>Note</u>: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. 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 that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR
- rationale/data support no risk adjustment/ stratification.

2b5. 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.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

2d. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:
2d1. 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
2d2.the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.
(*if not conducted or results not adequate, justification must be submitted and accepted)*

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).
11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.
12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.
13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.
14. Risk factors that influence outcomes should not be specified as exclusions.
15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.
16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE
Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation. If different data sources are used for different components in the composite, indicate the component after the checkbox.*)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☐ abstracted from paper record	☐ abstracted from paper record
☐ administrative claims	☐ administrative claims
☒ clinical database/registry	☒ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: **Click here to describe**	☐ other: **Click here to describe**

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).
STS Adult Cardiac Surgery Database Version 2.73

1.3. What are the dates of the data used in testing?
July 2011 – June 2014

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☐ individual clinician	☐ individual clinician
☒ group/practice	☒ group/practice
☒ hospital/facility/agency	☒ hospital/facility/agency
☐ health plan	☐ health plan
☐ other: **Click here to describe**	☐ other: **Click here to describe**

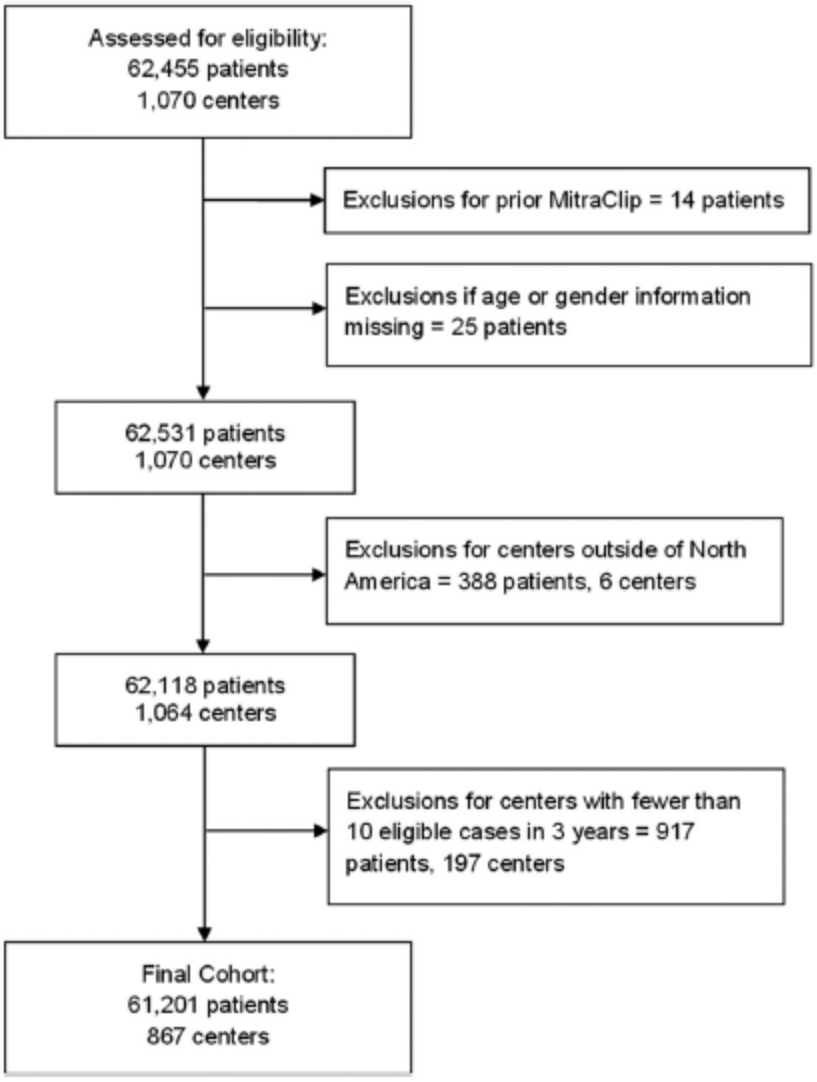
1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

The measure was developed and tested using STS Adult Cardiac Surgery database data from 867 participants for patients undergoing mitral valve repair/replacement (MVRR) during July 2011 – June 2014. Only participants with at least 10 eligible records during this period were included in the hierarchical model for estimating composite scores. The table below summarizes the distribution of participant-specific denominators (number of eligible patients) and participant-specific mortality and morbidity rates.

Stat	**N (Denominator)**	**% Mortality**	**% Morbidity**
N	867	867	867
Mean	71	3.8	19.0
STD	111	4.1	10.1
IQR	58	5.6	12.1
0%	10	0.0	0.0
10%	13	0.0	7.7
20%	18	0.0	10.7
30%	24	0.8	13.6
40%	30	2.0	15.5
50%	38	2.9	17.8
60%	50	3.9	20.0
70%	67	4.9	22.8
80%	95	6.3	25.9
90%	144	9.1	32.1
100%	1932	30.0	71.0

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

For assessing risk model discrimination and calibration, the sample included 62,118 patient operation records from 1,064 STS participants. For estimating composite scores, the sample was limited to participants with at least 10 eligible cases over the 3-year study period (i.e., 61,201 patients, 867 centers). The figure below summarizes the number of patients and centers included.



The table below summarizes the baseline characteristics of patients who were used for assessing risk model discrimination and calibration for the composite's two outcome measures.

	Effects	Overall N=62,118	MV Repair N=35,423	MV Replacement N=26,695
Age	Median (IQR)	64.0 (54.0, 73.0)	63.0 (54.0, 72.0)	65.0 (54.0, 74.0)
	Missing	0 (0.0%)	0 (0.0%)	0 (0.0%)
Age (years)	<55	15,813 (25.5%)	9,050 (25.5%)	6,763 (25.3%)
	>=55 and <65	16,241 (26.1%)	10,002 (28.2%)	6,239 (23.4%)
	>=65 and <75	17,058 (27.5%)	9,730 (27.5%)	7,328 (27.5%)
	>=75	13,006 (20.9%)	6,641 (18.7%)	6,365 (23.8%)
Gender	Male	31,040 (50.0%)	20,207 (57.0%)	10,833 (40.6%)
	Female	31,078 (50.0%)	15,216 (43.0%)	15,862 (59.4%)
Race	Caucasian	49,621 (79.9%)	29,656 (83.7%)	19,965 (74.8%)
	Black	6,085 (9.8%)	2,800 (7.9%)	3,285 (12.3%)
	Asian	1,571 (2.5%)	838 (2.4%)	733 (2.7%)
	Native American	216 (0.3%)	75 (0.2%)	141 (0.5%)
	Multiple Races	3,154 (5.1%)	1,325 (3.7%)	1,829 (6.9%)
	Other	1,149 (1.8%)	545 (1.5%)	604 (2.3%)
	Missing	322 (0.5%)	184 (0.5%)	138 (0.5%)
Ethnicity	Non-Hispanic	58,804 (94.7%)	34,051 (96.1%)	24,753 (92.7%)

	Effects	Overall N=62,118	MV Repair N=35,423	MV Replacement N=26,695
	Hispanic	2,990 (4.8%)	1,195 (3.4%)	1,795 (6.7%)
	Missing	324 (0.5%)	177 (0.5%)	147 (0.6%)
Body Surface Area (m)	<1.5	2,546 (4.1%)	1,172 (3.3%)	1,374 (5.1%)
	>=1.5 and <1.75	14,991 (24.1%)	7,705 (21.8%)	7,286 (27.3%)
	>=1.75 and <2	22,674 (36.5%)	12,807 (36.2%)	9,867 (37.0%)
	>=2	21,800 (35.1%)	13,682 (38.6%)	8,118 (30.4%)
	Missing	107 (0.2%)	57 (0.2%)	50 (0.2%)
Body Mass Index (kg/m)	<25	22,949 (36.9%)	13,405 (37.8%)	9,544 (35.8%)
	>=25 and <30	21,562 (34.7%)	13,097 (37.0%)	8,465 (31.7%)
	>=30 and <35	10,466 (16.8%)	5,691 (16.1%)	4,775 (17.9%)
	>=35	6,938 (11.2%)	3,129 (8.8%)	3,809 (14.3%)
	Missing	203 (0.3%)	101 (0.3%)	102 (0.4%)
Diabetes	No Diabetes	50,891 (81.9%)	30,670 (86.6%)	20,221 (75.7%)
	Diabetes - Noninsulin	7,722 (12.4%)	3,512 (9.9%)	4,210 (15.8%)
	Diabetes - Insulin	3,340 (5.4%)	1,146 (3.2%)	2,194 (8.2%)
	Diabetes - Missing Treatment	14 (0.0%)	5 (0.0%)	9 (0.0%)
	Missing	151 (0.2%)	90 (0.3%)	61 (0.2%)
Hypertension	No	20,249 (32.6%)	12,864 (36.3%)	7,385 (27.7%)
	Yes	41,739 (67.2%)	22,485 (63.5%)	19,254 (72.1%)
	Missing	130 (0.2%)	74 (0.2%)	56 (0.2%)
Dyslipidemia	No	27,855 (44.8%)	16,780 (47.4%)	11,075 (41.5%)
	Yes	34,128 (54.9%)	18,572 (52.4%)	15,556 (58.3%)
	Missing	135 (0.2%)	71 (0.2%)	64 (0.2%)
Cigarette Smoking	Never-smoker	49,371 (79.5%)	29,705 (83.9%)	19,666 (73.7%)
	Past Smoker	5,548 (8.9%)	2,724 (7.7%)	2,824 (10.6%)
	Current Smoker	7,062 (11.4%)	2,924 (8.3%)	4,138 (15.5%)
	Missing	137 (0.2%)	70 (0.2%)	67 (0.3%)
Chronic Lung Disease (CLD)	None	47,014 (75.7%)	28,690 (81.0%)	18,324 (68.6%)
	Mild	8,159 (13.1%)	3,902 (11.0%)	4,257 (15.9%)
	Moderate	3,820 (6.1%)	1,598 (4.5%)	2,222 (8.3%)
	Severe	2,864 (4.6%)	1,075 (3.0%)	1,789 (6.7%)
	Missing	261 (0.4%)	158 (0.4%)	103 (0.4%)
Peripheral Vascular Disease (PVD)	No	58,184 (93.7%)	33,772 (95.3%)	24,412 (91.4%)
	Yes	3,750 (6.0%)	1,547 (4.4%)	2,203 (8.3%)
	Missing	184 (0.3%)	104 (0.3%)	80 (0.3%)
Cerebrovascular Disease (CVD)	No	53,980 (86.9%)	32,350 (91.3%)	21,630 (81.0%)
	Yes	7,972 (12.8%)	2,982 (8.4%)	4,990 (18.7%)
	Missing	166 (0.3%)	91 (0.3%)	75 (0.3%)
Cerebrovascular Accident (CVA)	No CVA	56,735 (91.3%)	33,579 (94.8%)	23,156 (86.7%)
	Remote CVA (> 2 weeks)	4,334 (7.0%)	1,539 (4.3%)	2,795 (10.5%)
	Recent CVA (< 2 weeks)	822 (1.3%)	183 (0.5%)	639 (2.4%)
	CVA - Missing Timing	55 (0.1%)	26 (0.1%)	29 (0.1%)
	Missing	172 (0.3%)	96 (0.3%)	76 (0.3%)
Endocarditis	No Endocarditis	55,777 (89.8%)	33,391 (94.3%)	22,386 (83.9%)
	Treated Endocarditis	2,895 (4.7%)	1,246 (3.5%)	1,649 (6.2%)

	Effects	Overall N=62,118	MV Repair N=35,423	MV Replacement N=26,695
	Active Endocarditis	3,265 (5.3%)	687 (1.9%)	2,578 (9.7%)
	Endocarditis - Missing Type	17 (0.0%)	6 (0.0%)	11 (0.0%)
	Missing	164 (0.3%)	93 (0.3%)	71 (0.3%)
Renal Function	Creatinine <1 mg/dL	30,807 (49.6%)	18,349 (51.8%)	12,458 (46.7%)
	Creatinine 1-1.5 mg/dL	23,769 (38.3%)	14,021 (39.6%)	9,748 (36.5%)
	Creatinine 1.5-2 mg/dL	3,748 (6.0%)	1,680 (4.7%)	2,068 (7.7%)
	Creatinine 2-2.5 mg/dL	883 (1.4%)	319 (0.9%)	564 (2.1%)
	Creatinine >2.5 mg/dL	624 (1.0%)	212 (0.6%)	412 (1.5%)
	Dialysis	1,963 (3.2%)	661 (1.9%)	1,302 (4.9%)
	Missing	324 (0.5%)	181 (0.5%)	143 (0.5%)
Immunosuppressive Treatment	No	59,505 (95.8%)	34,311 (96.9%)	25,194 (94.4%)
	Yes	2,396 (3.9%)	983 (2.8%)	1,413 (5.3%)
	Missing	217 (0.3%)	129 (0.4%)	88 (0.3%)
Previous Coronary Artery Bypass Surgery	No	57,804 (93.1%)	33,919 (95.8%)	23,885 (89.5%)
	Yes	4,178 (6.7%)	1,430 (4.0%)	2,748 (10.3%)
	Missing	136 (0.2%)	74 (0.2%)	62 (0.2%)
Previous Valve Surgery	No	54,273 (87.4%)	34,303 (96.8%)	19,970 (74.8%)
	Yes	7,720 (12.4%)	1,049 (3.0%)	6,671 (25.0%)
	Missing	125 (0.2%)	71 (0.2%)	54 (0.2%)
Previous Other Cardiac Surgery	No	58,766 (94.6%)	33,985 (95.9%)	24,781 (92.8%)
	Yes	3,134 (5.0%)	1,344 (3.8%)	1,790 (6.7%)
	Missing	218 (0.4%)	94 (0.3%)	124 (0.5%)
Number of Previous CV Surgeries	No Previous CV surgery	51,613 (83.1%)	32,805 (92.6%)	18,808 (70.5%)
	1 Prior CV surgery	9,007 (14.5%)	2,329 (6.6%)	6,678 (25.0%)
	2 Or More Prior CV Surgeries	1,460 (2.4%)	276 (0.8%)	1,184 (4.4%)
	Missing	38 (0.1%)	13 (0.0%)	25 (0.1%)
Prior PCI	No PCI	56,546 (91.0%)	32,771 (92.5%)	23,775 (89.1%)
	PCI - within 6 hours	67 (0.1%)	17 (0.0%)	50 (0.2%)
	PCI - not within 6 hours	5,352 (8.6%)	2,555 (7.2%)	2,797 (10.5%)
	PCI - missing timing	17 (0.0%)	7 (0.0%)	10 (0.0%)
	Missing	136 (0.2%)	73 (0.2%)	63 (0.2%)
Acuity Status	Elective	46,810 (75.4%)	29,603 (83.6%)	17,207 (64.5%)
	Urgent	14,312 (23.0%)	5,613 (15.8%)	8,699 (32.6%)
	Emergent	892 (1.4%)	174 (0.5%)	718 (2.7%)
	Emergent Salvage	44 (0.1%)	1 (0.0%)	43 (0.2%)
	Missing	60 (0.1%)	32 (0.1%)	28 (0.1%)
Myocardial Infarction	No Prior MI	55,877 (90.0%)	32,751 (92.5%)	23,126 (86.6%)
	MI >21 days	5,114 (8.2%)	2,323 (6.6%)	2,791 (10.5%)
	MI 8-21 days	361 (0.6%)	118 (0.3%)	243 (0.9%)
	MI 1-7 days	479 (0.8%)	131 (0.4%)	348 (1.3%)

	Effects	Overall N=62,118	MV Repair N=35,423	MV Replacement N=26,695
	MI 6-24 hrs	75 (0.1%)	9 (0.0%)	66 (0.2%)
	MI <= 6 hrs	37 (0.1%)	4 (0.0%)	33 (0.1%)
	MI - Missing Timing	23 (0.0%)	9 (0.0%)	14 (0.1%)
	Missing	152 (0.2%)	78 (0.2%)	74 (0.3%)
Angina	No Symptoms or Angina	28,961 (46.6%)	17,542 (49.5%)	11,419 (42.8%)
	Symptoms Unlikely to be Ischemia	27,789 (44.7%)	15,163 (42.8%)	12,626 (47.3%)
	Stable Angina	2,390 (3.8%)	1,409 (4.0%)	981 (3.7%)
	Unstable Angina	1,807 (2.9%)	854 (2.4%)	953 (3.6%)
	Non-ST Elevation MI (Non-STEMI)	539 (0.9%)	164 (0.5%)	375 (1.4%)
	ST Elevation MI (STEMI)	169 (0.3%)	21 (0.1%)	148 (0.6%)
	Missing	463 (0.7%)	270 (0.8%)	193 (0.7%)
Cardiogenic Shock	No	61,001 (98.2%)	35,126 (99.2%)	25,875 (96.9%)
	Yes	968 (1.6%)	215 (0.6%)	753 (2.8%)
	Missing	149 (0.2%)	82 (0.2%)	67 (0.3%)
Resuscitation	No	61,646 (99.2%)	35,225 (99.4%)	26,421 (99.0%)
	Yes	319 (0.5%)	114 (0.3%)	205 (0.8%)
	Missing	153 (0.2%)	84 (0.2%)	69 (0.3%)
Arrhythmia	AFib/Flutter	19,517 (31.4%)	9,911 (28.0%)	9,606 (36.0%)
	Heart Block	94 (0.2%)	25 (0.1%)	69 (0.3%)
	Sustained VT/VF	372 (0.6%)	200 (0.6%)	172 (0.6%)
	Multiple Types	491 (0.8%)	221 (0.6%)	270 (1.0%)
	Arrhythmia - Other	163 (0.3%)	86 (0.2%)	77 (0.3%)
	Arrhythmia - Missing Type	9 (0.0%)	4 (0.0%)	5 (0.0%)
	Missing	41,472 (66.8%)	24,976 (70.5%)	16,496 (61.8%)
Preop IABP	No	60,801 (97.9%)	34,977 (98.7%)	25,824 (96.7%)
	Yes	1,168 (1.9%)	380 (1.1%)	788 (3.0%)
	Missing	149 (0.2%)	66 (0.2%)	83 (0.3%)
Congestive Heart Failure	No	30,550 (49.2%)	19,298 (54.5%)	11,252 (42.2%)
	Yes	31,436 (50.6%)	16,049 (45.3%)	15,387 (57.6%)
	Missing	132 (0.2%)	76 (0.2%)	56 (0.2%)
NYHA Classification in Patients with CHF	I	1,616 (5.1%)	1,103 (6.9%)	513 (3.3%)
	II	8,724 (27.8%)	5,462 (34.0%)	3,262 (21.2%)
	III	13,562 (43.1%)	6,551 (40.8%)	7,011 (45.6%)
	IV	6,860 (21.8%)	2,651 (16.5%)	4,209 (27.4%)
	Missing	674 (2.1%)	282 (1.8%)	392 (2.5%)
Number of Diseased Coronary Vessels	None	51,522 (82.9%)	30,501 (86.1%)	21,021 (78.7%)
	One	4,722 (7.6%)	2,314 (6.5%)	2,408 (9.0%)
	Two	2,038 (3.3%)	952 (2.7%)	1,086 (4.1%)
	Three	2,581 (4.2%)	1,048 (3.0%)	1,533 (5.7%)
	Missing	1,255 (2.0%)	608 (1.7%)	647 (2.4%)
Left Main Disease > 50%	No	60,186 (96.9%)	34,527 (97.5%)	25,659 (96.1%)
	Yes	726 (1.2%)	303 (0.9%)	423 (1.6%)
	Missing	1,206 (1.9%)	593 (1.7%)	613 (2.3%)
Ejection Fraction (%)	<25	894 (1.4%)	551 (1.6%)	343 (1.3%)
	>=25 and <35	2,499 (4.0%)	1,378 (3.9%)	1,121 (4.2%)
	>=35 and <45	4,754 (7.7%)	2,475 (7.0%)	2,279 (8.5%)

	Effects	Overall N=62,118	MV Repair N=35,423	MV Replacement N=26,695
	>=45 and <55	9,990 (16.1%)	5,387 (15.2%)	4,603 (17.2%)
	>=55	42,269 (68.0%)	24,780 (70.0%)	17,489 (65.5%)
	Missing	1,712 (2.8%)	852 (2.4%)	860 (3.2%)
Aortic Stenosis	No	60,268 (97.0%)	34,773 (98.2%)	25,495 (95.5%)
	Yes	1,484 (2.4%)	434 (1.2%)	1,050 (3.9%)
	Missing	366 (0.6%)	216 (0.6%)	150 (0.6%)
Mitral Stenosis	No	51,857 (83.5%)	33,837 (95.5%)	18,020 (67.5%)
	Yes	8,908 (14.3%)	754 (2.1%)	8,154 (30.5%)
	Missing	1,353 (2.2%)	832 (2.3%)	521 (2.0%)
Tricuspid Stenosis	No	60,365 (97.2%)	34,531 (97.5%)	25,834 (96.8%)
	Yes	292 (0.5%)	116 (0.3%)	176 (0.7%)
	Missing	1,461 (2.4%)	776 (2.2%)	685 (2.6%)
Pulmonic Stenosis	No	61,588 (99.1%)	35,137 (99.2%)	26,451 (99.1%)
	Yes	116 (0.2%)	53 (0.1%)	63 (0.2%)
	Missing	414 (0.7%)	233 (0.7%)	181 (0.7%)
Aortic Insufficiency	None	45,116 (72.6%)	26,145 (73.8%)	18,971 (71.1%)
	Trivial	7,480 (12.0%)	4,456 (12.6%)	3,024 (11.3%)
	Mild	7,451 (12.0%)	3,865 (10.9%)	3,586 (13.4%)
	Moderate	1,885 (3.0%)	889 (2.5%)	996 (3.7%)
	Severe	129 (0.2%)	47 (0.1%)	82 (0.3%)
	Missing	57 (0.1%)	21 (0.1%)	36 (0.1%)
Mitral Insufficiency	None	1,752 (2.8%)	479 (1.4%)	1,273 (4.8%)
	Trivial	974 (1.6%)	141 (0.4%)	833 (3.1%)
	Mild	2,337 (3.8%)	387 (1.1%)	1,950 (7.3%)
	Moderate	6,692 (10.8%)	2,930 (8.3%)	3,762 (14.1%)
	Severe	50,027 (80.5%)	31,332 (88.5%)	18,695 (70.0%)
	Missing	336 (0.5%)	154 (0.4%)	182 (0.7%)
Tricuspid Insufficiency	None	22,173 (35.7%)	12,864 (36.3%)	9,309 (34.9%)
	Trivial	8,136 (13.1%)	5,523 (15.6%)	2,613 (9.8%)
	Mild	15,020 (24.2%)	8,895 (25.1%)	6,125 (22.9%)
	Moderate	10,843 (17.5%)	5,446 (15.4%)	5,397 (20.2%)
	Severe	5,724 (9.2%)	2,581 (7.3%)	3,143 (11.8%)
	Missing	222 (0.4%)	114 (0.3%)	108 (0.4%)
Pulmonic Insufficiency	None	46,161 (74.3%)	26,070 (73.6%)	20,091 (75.3%)
	Trivial	9,653 (15.5%)	6,017 (17.0%)	3,636 (13.6%)
	Mild	5,263 (8.5%)	2,877 (8.1%)	2,386 (8.9%)
	Moderate	947 (1.5%)	422 (1.2%)	525 (2.0%)
	Severe	63 (0.1%)	25 (0.1%)	38 (0.1%)
	Missing	31 (0.0%)	12 (0.0%)	19 (0.1%)

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

For assessing risk model discrimination and calibration, we used data from 62,118 eligible patients at 1,064 participants undergoing MVRR during July 2011 – June 2014.

For estimating participant-specific composite scores, the analysis was restricted to data from STS participants with at least 10 eligible cases during July 2011 – June 2014 (N = 61,201 patient records, 867 participants).

For assessing the consistency of results over time, we re-estimated composite scores using data from July 2012 – June 2015 (N = 63,516 patient records, 873 participants). This analysis included all participants with at least 10 eligible cases during July 2012 – June 2015.

To ensure adequate statistical precision, the STS plans to report composite scores only for participants with at least 36 eligible cases during the 3-year measurement window. Thus, some of the analyses in this submission are limited to participants with at least 36 eligible cases.

2a2. RELIABILITY TESTING

2a2.1. What level of reliability testing was conducted?

Note*: Current guidance for composite measure evaluation states that reliability must be demonstrated for the composite performance measure score.*

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. Describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Reliability is conventionally defined as the proportion of variation in a performance measure that is due to true between-hospital differences (i.e., signal) as opposed to random statistical fluctuations (i.e., noise). A mathematically equivalent definition is the squared correlation between a measurement and the true value. This quantity cannot be calculated directly because the "true" composite measure values are unknown, but may be estimated, as described below.

Calculation Details

Let θ_j denote the true unknown composite measure value for the j-th of J participants. Before estimating reliability, the numeric value of θ_j was estimated for each participant under the assumed hierarchical model. Estimation was done using Markov Chain Monte Carlo (MCMC) simulations and involved the following steps:

1. For each *j,* we randomly generated a large number (*N*) of possible numeric values of θ_j by sampling from the Bayesian posterior probability distribution of θ_j via MCMC sampling. Let $\theta_j^{(i)}$ denote the i-th of these *N* randomly sampled numerical values for the j-th participant.
2. For each j, the posterior mean $\hat{\theta}_j$ of θ_j was calculated as the arithmetic average of the randomly sampled values $\theta_j^{(1)}, \ldots, \theta_j^{(N)}$; in other words $\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N} \theta_j^{(i)}$.

Our reliability measure was defined as the squared correlation between the set of hospital-specific estimates $\hat{\theta}_1, \ldots, \hat{\theta}_J$ and the corresponding unknown true values $\theta_1, \ldots, \theta_J$. Let ρ^2 denote the unknown true squared correlation of interest and let $\hat{\rho}^2$ denote an estimate of this quantity. The estimate was calculated as

$$\hat{\rho}^2 = \frac{1}{N}\sum_{i=1}^{N} \rho_{(i)}^2$$

where

$$\rho_{(i)}^2 = \frac{\left[\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)\left(\hat{\theta}_j - \bar{\theta}\right)\right]^2}{\sum_{j=1}^{J}\left(\theta_j^{(i)} - \bar{\theta}^{(i)}\right)^2 \sum_{j=1}^{J}\left(\hat{\theta}_j - \bar{\theta}\right)^2}, \quad \bar{\theta} = \frac{1}{JN}\sum_{j=1}^{J}\sum_{i=1}^{N} \theta_j^{(i)} \quad \text{and} \quad \bar{\theta}^{(i)} = \frac{1}{J}\sum_{j=1}^{J} \theta_j^{(i)}.$$

A 95% Bayesian probability interval for ρ^2 was obtained calculating the 2.5th and 97.5th percentiles of the set of numbers $\rho_{(1)}^2, \ldots, \rho_{(N)}^2$.

2a2.3. What were the statistical results from reliability testing? (e.g., *percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

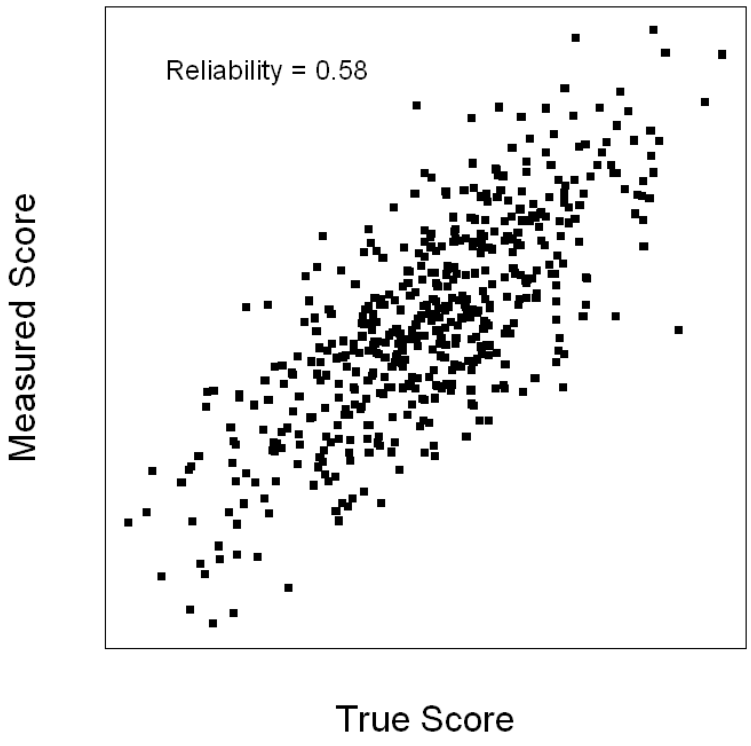
The estimated reliability of the STS MVRR composite measure using 3 years of data in participants with at least 36 total cases was 0.58 (95% CrI, 0.52 to 0.64), as outlined in the Table below. For comparison, the reliability of the STS isolated CABG composite score was 0.77 (95% CrI, 0.74 to 0.80) using 1 year of data in 2013. Using 3 years of data from 2011 to 2013, the reliability of the STS AVR composite measure was 0.52 (95% CrI, 0.47 to 0.57), and the AVR+CABG measure was 0.50 (95% CrI, 0.45 to 0.54)

Time Span	Participants Included (No.)	Patients Included (No.)	Reliability $\widehat{\rho}^2$ (95% PrI)
3 years	867	61,201	0.47 (0.42–0.52)
3 years and Participants with at least			
25 cases	597	56,799	0.55 (0.49–0.60)
36 cases	462	52,841	0.58 (0.52–0.64)
50 cases	349	48,076	0.61 (0.54–0.66)
100 cases	165	35,198	0.69 (0.62–0.76)

Based in part on these results, we selected a threshold of 1 mitral case per month, or 36 cases over 3 years, as a minimum threshold for receiving a site-specific STS MVRR composite score. This resulted in a reliability of 0.58 but reduced the number of programs eligible to receive a score from 867 to 462. A higher volume threshold would have yielded even higher reliability but at the cost of further reducing the number of programs eligible to receive a score.

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

To interpret the results, we created a figure illustrating the accuracy of the measured scores when the true reliability = 0.58. Because the true score for the composite measure is unknown, we used simulated data with formula Measured Score$_i$=True Score$_i$ + e_i where $i = 1,2,\ldots,462$ indicates the 462 participants and where True Score$_i$ and e_i both follow normal distributions. The standard deviations of the normal distributions were chosen such that the measure has a reliability of 0.58.



2b2. VALIDITY TESTING

***Note**: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b2.1. What level of validity testing was conducted?

☒ **Composite performance measure score**

- ☐ **Empirical validity testing**
- ☐ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)
- ☐ **Systematic assessment of content validity**

☐ **Validity testing for component measures** *(check all that apply)*

***Note**: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

- ☐ **Endorsed (or submitted) as individual performance measures**
- ☐ **Critical data elements** (*data element validity must address ALL critical data elements*)
- ☐ **Empirical validity testing of the component measure score(s)**
- ☐ **Systematic assessment of face validity of component measure score(s) as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

The tests on validity used the concept of performance categories to be more formally introduced in 2b5: Participants were labeled as having higher-than-expected performance if the 95% credible interval surrounding a participant's

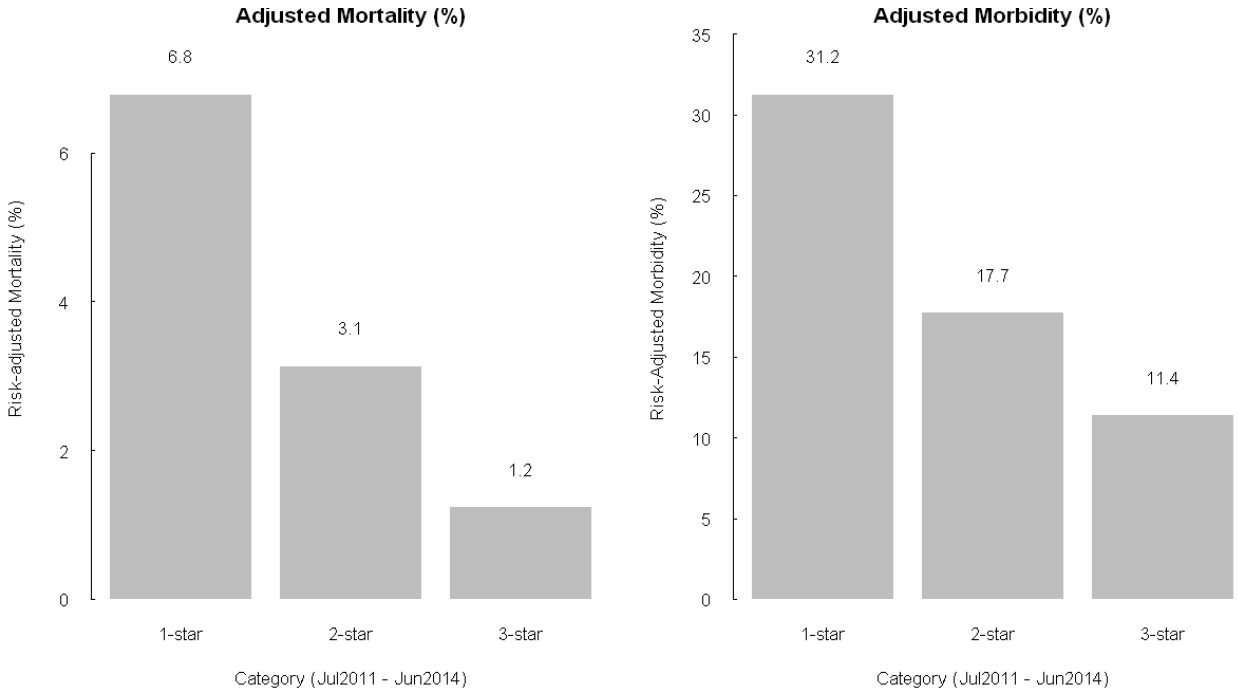
composite score fell entirely above the overall STS average composite score. Participants were labeled as having lower-than-expected performance if the 95% credible interval surrounding a participant's composite score fell entirely below the overall STS average composite score. Participants were labeled as higher-than-expected performance (3 stars), lower-than-expected performance (1 star), and indistinguishable from the average or as-expected performance (2 stars).

We compared risk-adjusted mortality and morbidity rates across the three performance groups. The measure has good face value if the three groups have different proportions as expected.

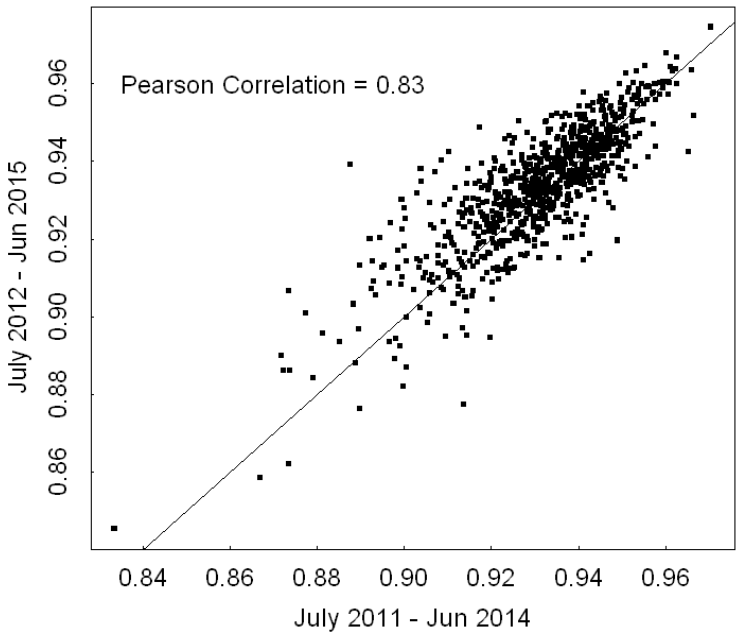
In addition, we assessed the extent to which a participant's composite score remains stable across two consecutive overlapping reporting periods. This analysis was restricted to 835 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Compared to participants receiving 1 star, those with 3 stars had lower risk-adjusted mortality (1.2% vs. 6.8%) and lower risk-adjusted morbidity (11.4% vs. 31.2%) during July 2011 – June 2014. Thus, differences in performance were clinically meaningful as well as statistically significant. STS participants deemed better by the composite scores have (on average) higher performance during the same time window on each individual domain of the composite measure.

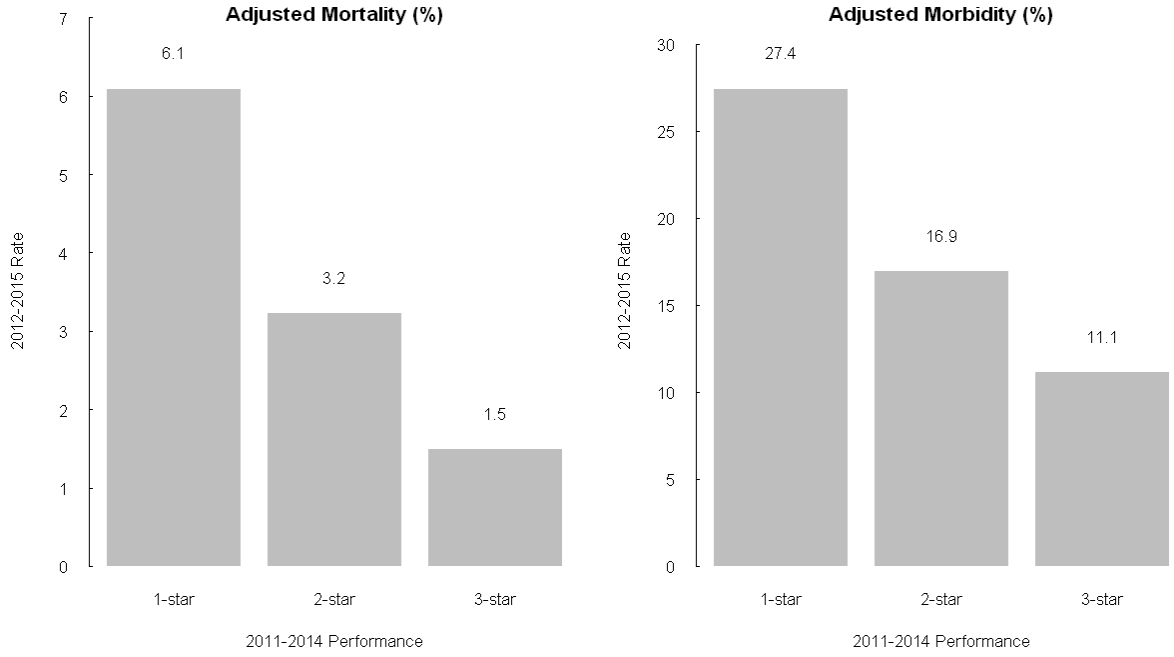


Stability of the composite measure over time was assessed in 835 participants who participated and had at least 10 eligible cases in each of two consecutive reporting periods: July 2011 – June 2014 and July 2012 – June 2015.



The Pearson correlation between the composite score calculated in the earlier and later time period was 0.83.

Using data from July 2012 – June 2015, we compared risk-adjusted mortality and morbidity rates across participants categories based on their composite measure performance in July 2011 – June 2014. Compared to 1-star participants, those with 3 stars had lower risk-adjusted mortality (1.5% versus 6.1%) and risk-adjusted morbidity (11.1% versus 27.4%) during July 2012 – June 2015.



2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i.*e., what do the results mean and what are the norms for the test conducted?*)

The test results show that the composite measure behaves as expected and that results are reasonably consistent across two consecutive overlapping time periods. These results support the validity of the composite measure as a quality measure for mitral valve replacement/repair procedures.

2b3. EXCLUSIONS ANALYSIS

Note: *Applies to the composite performance measure, as well all component measures unless they are already endorsed or are being submitted for individual endorsement.*

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

N/A

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

N/A

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

N/A

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

Note: *Applies to all outcome or resource use component measures, unless already endorsed or are being submitted for individual endorsement.*

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☒ **Endorsed (or submitted) as individual performance measures**
☐ **No risk adjustment or stratification**
☒ **Statistical risk model**
☐ **Stratification by risk categories**
☐ **Other, Click here to enter description**

2b4.2. If an outcome or resource use component measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

N/A

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

To adjust for case mix in the STS MVRR Composite Score [1], the published 2008 STS isolated valve model [2] was modified and re-estimated in the current study population. The main reason for modifying the model was to be able to calculate predicted risk estimates for patients in the current study population who did not meet inclusion/exclusion criteria for the existing 2008 STS isolated valve model. In addition, although the existing STS models predict the endpoints of “operative mortality” and “operative mortality or major morbidity” there is no existing model for predicting “major morbidity” as defined in the current study. In the future, as STS risk models are revised over time, the composite measure will be calculated with the most up to date STS risk model for the MVRR population.

Except where noted below, covariates for the modified operative mortality model were identical to the STS 2008 operative mortality model and covariates for the new major morbidity model were identical to the STS 2008 operative mortality or major morbidity model. For each of the two models, the list of covariates was modified as follows.

- Adjust for concomitant tricuspid repair. The STS 2008 models excluded patients undergoing a concomitant tricuspid procedure. Because the current study included patients undergoing concomitant tricuspid repair, an indicator variable for tricuspid repair was included.
- Adjust for tricuspid insufficiency using categories none or mild, moderate, and severe. The 2008 models included indicators of at least moderate tricuspid insufficiency. Because of the inclusion of operations with concurrent TV repair procedures, the surgeon panel felt it was necessary to more finely adjust the degrees of tricuspid insufficiency. The modified models include separate indicator variables for moderate tricuspid insufficiency and severe tricuspid insufficiency.
- Adjust for infectious endocarditis using categories active, treated, and none. The 2008 models include an indicator for active infections endocarditis but not for treated infectious endocarditis. The modified models include separate indicator variables for treated infectious endocarditis and active infectious endocarditis.

Considerations for adjusting for tricuspid repair

It is a generally accepted principle not to use what may be discretionary procedural decisions (e.g., whether or not to add a tricuspid valve repair) in profiling models. However, as discussed in the main article, there is accumulating evidence of the potential longitudinal merits of concomitant TVr. and the surgeon panel wanted to avoid discouraging the performance of this procedure by failing to account for its increased inherent risk of morbidity. Furthermore, the panel felt that the need to perform TVr may be a proxy for more advanced disease that may not be captured perfectly in the current STS data collection form.

References

1. Badhwar V, Rankin JS, He X, Jacobs JP, Gammie JS, Furnary AP, Fazzalari FL, Han J, O'Brien SM, Shahian DM. The Society of Thoracic Surgeons Mitral Repair/Replacement Composite Score: A Report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg. 2016 Jun;101(6):2265-71. doi: 10.1016/j.athoracsur.2015.11.049. Epub 2015 Dec 28.
2. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.

2b4.4. What were the statistical results of the analyses used to select risk factors?

Estimated odds ratios from the modified STS 2008 models are summarized in the table below.

	Morbidity		Mortality	
Effect	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
Effects that do not interact with MV repair/replacements				
Preoperative atrial fibrillation	1.14 (1.08, 1.20)	<.0001	1.22 (1.09, 1.36)	0.0005
Race (v. others)				
Black	1.25 (1.15, 1.35)	<.0001	NA	.
Hispanic	1.23 (1.10, 1.39)	0.0003	NA	.
CVD (v. no)				
CVD with CVA	1.17 (1.08, 1.27)	0.0002	NA	.
CVD without CVA	0.97 (0.88, 1.08)	0.6072	NA	.
Number Diseased Vessels (3 v. 2, 2 v. 1/0)	1.06 (1.00, 1.12)	0.0681	NA	.
Pre-op IABP or inotrope	2.20 (1.93, 2.52)	<.0001	1.35 (1.10, 1.66)	0.0040
Hypertension	1.13 (1.06, 1.20)	<.0001	1.11 (0.98, 1.27)	0.1115
Immunosuppressive treatment	1.26 (1.13, 1.41)	<.0001	1.60 (1.33, 1.93)	<.0001
Peripheral vascular disease	1.24 (1.14, 1.35)	<.0001	1.31 (1.13, 1.52)	0.0004

Aortic stenosis	1.12 (0.99, 1.27)	0.0714	NA	.
MI <21 days	1.40 (1.19, 1.63)	<.0001	1.77 (1.41, 2.21)	<.0001
Shock	2.52 (2.08, 3.06)	<.0001	1.84 (1.38, 2.47)	<.0001
Number of previous operations (v. 0)				
1 previous operation	1.35 (1.21, 1.50)	<.0001	1.69 (1.34, 2.15)	<.0001
2 or more previous operations	1.65 (1.39, 1.95)	<.0001	2.12 (1.53, 2.94)	<.0001
Urgent status (v. elective)	1.32 (1.23, 1.41)	<.0001	1.27 (1.12, 1.44)	0.0003
Active infections endocarditis	1.80 (1.63, 1.99)	<.0001	1.86 (1.54, 2.23)	<.0001
Treated infections endocarditis	1.07 (0.96, 1.19)	0.2471	0.96 (0.76, 1.22)	0.7427
Ejection fraction per 10-unit decrease	1.11 (1.07, 1.15)	<.0001	1.17 (1.09, 1.25)	<.0001
Creatinine per 1 unit increase	1.62 (1.53, 1.71)	<.0001	1.48 (1.36, 1.62)	<.0001
Body surface area, m^2				
1.6 v. 2.0 in male	1.26 (1.11, 1.44)	0.0004	1.64 (1.29, 2.09)	<.0001
1.8 v. 2.0 in male	1.05 (1.00, 1.10)	0.0418	1.19 (1.08, 1.30)	0.0002
2.2 v. 2.0 in male	1.09 (1.05, 1.13)	<.0001	0.99 (0.91, 1.06)	0.6966
1.6 v. 1.8 in female	1.07 (1.03, 1.12)	0.0017	1.25 (1.16, 1.35)	<.0001
2.0 v. 1.8 in female	1.04 (1.01, 1.08)	0.0074	0.98 (0.92, 1.04)	0.4325
2.2 v. 1.8 in female	1.21 (1.12, 1.31)	<.0001	1.17 (1.00, 1.37)	0.0500
Time trend (half year increase)	0.96 (0.95, 0.98)	<.0001	1.01 (0.98, 1.04)	0.6571
Left main disease	NA	.	1.11 (0.81, 1.53)	0.5091
Unstable angina (no MI < 8days)	NA	.	1.21 (0.94, 1.55)	0.1420
Mitral stenosis	NA	.	1.05 (0.91, 1.20)	0.5060
Moderate tricuspid insufficiency (v. no-mild)	1.13 (1.05, 1.20)	0.0003	1.17 (1.02, 1.33)	0.0243
Severe tricuspid insufficiency (v. no-mild)	1.23 (1.12, 1.35)	<.0001	1.46 (1.22, 1.75)	<.0001
Mitral valve repair (v. replacement)	0.56 (0.50, 0.62)	<.0001	0.40 (0.31, 0.53)	<.0001
Tricuspid valve repair (v. none)	1.36 (1.24, 1.49)	<.0001	0.99 (0.84, 1.18)	0.9474
Effects that interacts with procedure groups and were modeled separately for MV replacement and MV repairs				
In MV replacements				
Age				
60 v. 50 (no reoperations, non-emergent)	1.22 (1.17, 1.27)	<.0001	1.53 (1.40, 1.67)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.49 (1.37, 1.62)	<.0001	2.34 (1.95, 2.81)	<.0001
80 v. 50 (no reoperations, non-emergent)	1.80 (1.61, 2.01)	<.0001	3.69 (2.95, 4.63)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.08 (1.00, 1.17)	0.0517	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.53 (1.38, 1.69)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)				
Insulin diabetes	1.41 (1.27, 1.57)	<.0001	1.48 (1.25, 1.75)	<.0001
Non-insulin diabetes	1.10 (1.02, 1.20)	0.0174	1.14 (1.00, 1.31)	0.0587
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.15 (1.11, 1.18)	<.0001	1.20 (1.13, 1.28)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	2.59 (2.12, 3.15)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.32 (1.13, 1.53)	0.0004
Status (v. elective)				
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	2.38 (1.61, 3.49)	<.0001
Emergent+resuscitation/emergent salvage	2.83 (1.63, 4.89)	0.0002	5.91 (3.18, 10.98)	<.0001
In MV repairs				
Age				
60 v. 50 (no reoperations, non-emergent)	1.27 (1.21, 1.32)	<.0001	1.76 (1.57, 1.97)	<.0001
70 v. 50 (no reoperations, non-emergent)	1.60 (1.47, 1.75)	<.0001	3.09 (2.46, 3.88)	<.0001
80 v. 50 (no reoperations, non-emergent)	2.00 (1.78, 2.25)	<.0001	5.61 (4.19, 7.50)	<.0001
Congestive heart failure (v. no)				
CHF not NYHA IV	1.17 (1.07, 1.28)	0.0007	1.20 (1.06, 1.37)	0.0043
CHF NYHA IV	1.55 (1.36, 1.76)	<.0001	1.66 (1.40, 1.96)	<.0001
Diabetes (v. no)				
Non-insulin diabetes	1.10 (0.99, 1.21)	0.0690	1.14 (1.00, 1.31)	0.0587

Insulin diabetes	1.44 (1.24, 1.66)	<.0001	1.48 (1.25, 1.75)	<.0001
Chronic lung disease (severe v moderate, or moderate v none-mild)	1.15 (1.11, 1.18)	<.0001	1.26 (1.16, 1.38)	<.0001
Dialysis v. no dialysis & creatinine = 1.0	1.97 (1.75, 2.22)	<.0001	3.68 (2.44, 5.54)	<.0001
Female (at BSA=1.8) v. male (at BSA=2.0)	1.17 (1.11, 1.25)	<.0001	1.14 (0.93, 1.40)	0.2108
Status (v. elective)				
Emergent - no resuscitation	3.30 (2.55, 4.27)	<.0001	3.83 (1.87, 7.86)	0.0003
Emergent+resuscitation/Emergent Salvage	2.83 (1.63, 4.89)	0.0002	2.47 (0.10, 59.60)	0.5785

CHF = congestive heart failure; CVA = cerebrovascular accident (stroke); CVD = cardiovascular disease; EF = ejection fraction; IABP = intra-aortic balloon pump; MI = myocardial infarction; NA = variable not used in model and estimate not available; NYHA = New York Heart Association.

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

The modified models were assessed using data from 62,118 patients undergoing MVRR during July 2011 – June 2014.

Discrimination

To gauge discrimination, we calculated the c-statistics of both models. Bootstrapping was used to estimate and adjust for the "optimism" from estimating and evaluating the model on the same sample [1].

Calibration

The model fit was evaluated using 5-fold cross validation. The entire sample was randomly split into five equal sized groups. The calibration plot was created by following these steps:

1. One of the five groups was used as the testing sample
2. The other four groups were combined into the training sample
3. The revised model was estimated using the training sample
4. The expected probability of experience the event in the testing sample was calculated using the model estimated in step 3.
5. The expected probability (from step 4) and observed event rates were then compared in the testing sample and the calibration plot was created.

The above five steps were repeated five times so that each group was used as the testing sample once. In the end, we had five calibration plots for each model.

Reference

1. Harrell, F. E., Kerry L. Lee, and Daniel B. Mark. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in medicine 15 (1996): 361-387.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
if stratified, skip to 2b4.9

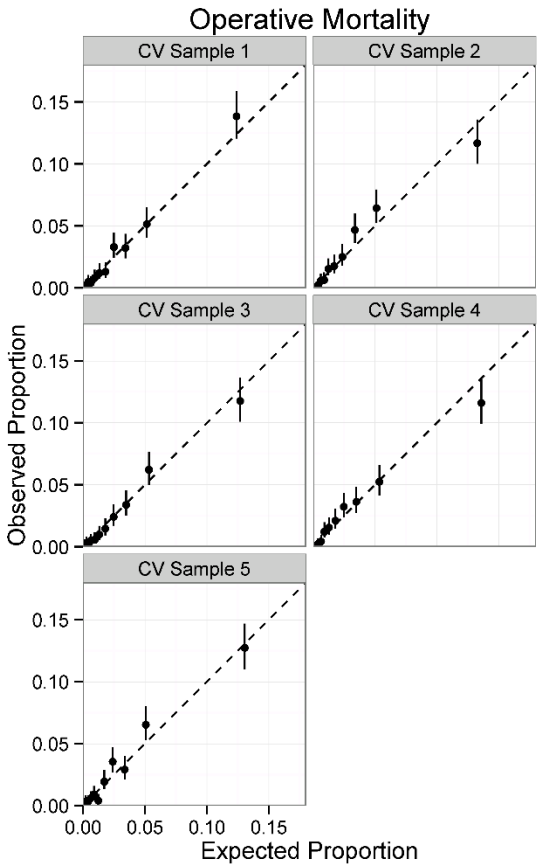
2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

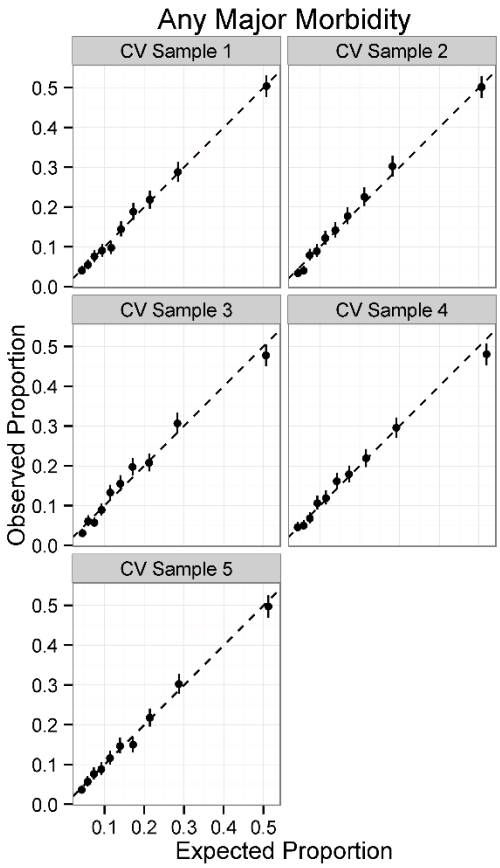
The bootstrap-adjusted estimated C-statistic was 0.746 for the morbidity model and 0.807 for the mortality model. These numbers were comparable to the STS 2008 valve models when evaluated using the same sample (0.745 and 0.807 for morbidity and mortality endpoints, respectively.)

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

N/A. The Hosmer-Lemeshow statistic was not calculated.

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:





2b4.9. Results of Risk Stratification Analysis:

N/A

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i*.e., what do the results mean and what are the norms for the test conducted?*)

The results demonstrated that the STS valve risk models are well calibrated and have good discrimination power. They are suitable for controlling differences in case-mix between centers.

***2b4.11. Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

***Note**: Applies to the composite performance measure.*

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

The degree of uncertainty surrounding an STS participant's composite measure estimate is indicated by calculating 95% Bayesian credible intervals (CI's) which are similar to conventional confidence intervals. Point estimates and CI's for an

individual STS participant are reported along with a comparison to various benchmarks based on the national sample. Benchmarks include the overall average STS composite score and several percentiles (minimum, 10th, 25th, 75th, 90th, maximum). In addition, the composite measure result is converted into categories labeled as 1 to 3 stars. An STS participant receives 2 stars if the Bayesian credible interval surrounding their composite score overlaps the overall STS average. This rating implies that the STS participant's performance was not statistically different from the overall STS national average. If the Bayesian CI falls entirely above the STS national average, the participant receives 3 stars (higher-than-expected performance). If the Bayesian CI falls entirely below the STS national average, the participant receives 1 star (lower-than-expected performance).

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Among participants with at least 36 cases over 3 years, around 85% of participants have received 2 stars, and the remaining participants have received either 1 or 3 stars.

Performance categories

July 2011 – June 2014

	All Participants	Participants N≥ 36
Category	**Number of Participants, %**	**Number of Participants, %**
1-star	23, 2.7%	16, 3.5%
2-star	795, 91.7%	397, 85.9%
3-star	49, 5.7%	49, 10.6%

July 2012 – June 2015

	All Participants	Participants N≥ 36
Category	**Number of Participants, %**	**Number of Participants, %**
1-star	19, 2.2%	16, 3.4%
2-star	798, 91.4%	400, 85.1%
3-star	56, 6.4%	54, 11.5%

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i.*e., what do the results mean in terms of statistical and meaningful differences?*)

The Bayesian methodology allows direct probability interpretation of the results. The identified differences in performance are both statistically significant and clinically meaningful. The surgeon panel and users are satisfied with the distribution of participants across performance categories.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS

Note: *Applies to all component measures, unless already endorsed or are being submitted for individual endorsement.* ***If only one set of specifications for each component, this section can be skipped.***

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the*

numerator). ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)

N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)

N/A

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i*.e., what do the results mean and what are the norms for the test conducted?*)

N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

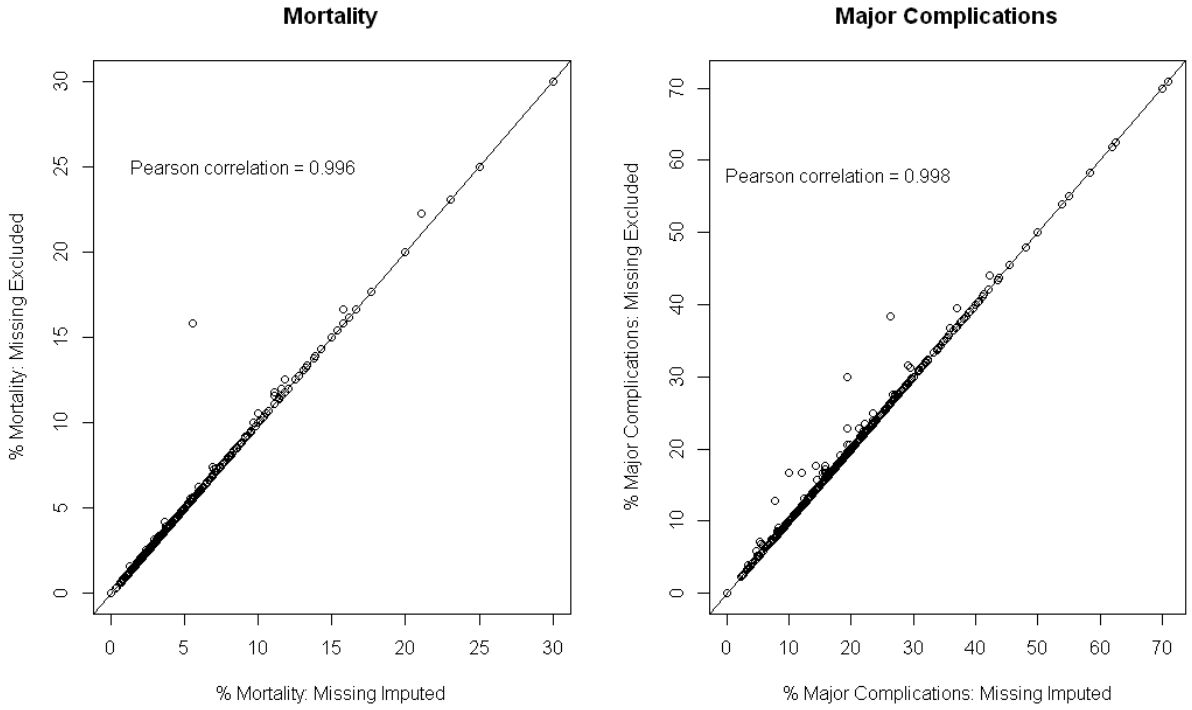
Missing data for risk model covariates was extremely rare: All model predictors had <5% missing and the majority had <1% missing. Missing data occurred in 0.5% of records for operative mortality and 0.4% of records for major complications. In the rare case of missing data, unknown values were imputed to the most common category of categorical variables and to the median or subgroup-specific median of continuous variables. Single imputation was used in the multivariable models consistent with the STS methodology used in the creation of the original STS model. More computationally intensive missing data strategies, such as multiple imputation, were not used for this analysis because of the low rate of missing data and because it would be impractical to implement them in combination with the computationally intensive Bayesian Markov Chain Monte Carlo (MCMC) procedure used for estimation of composite scores. Moreover, the use of multiple imputation has had negligible impact in previous STS analyses with similar low rates of missing data. For a comparison of single versus multiple imputation results in the development of the STS 2008 risk model, please see http://people.duke.edu/~obrie027/STS2008/. In that analysis, using multiple imputation did not appreciably widen the confidence intervals around model estimates. Moreover, any differences in point estimates were small relative to their standard error. Similar results have been found in a number of STS publication analyses.

A 30-day vital status category of "unknown" is available for those instances (e.g., homeless patients) in which the status of the patient cannot be ascertained despite good faith efforts to do so. In order to prevent excessive or inappropriate use of this vital status category, stringent new limitations on the use of this category were implemented in 2016, retroactive to 2015 data. This will further assure the accuracy of the operative mortality endpoint, which includes a small percentage of patients who die between hospital discharge and 30 days.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)

The overall frequency of missing data was 0.5% for operative mortality and 0.4% for major complications. The median participant-specific frequency of missing data was 0% (range 0% to 69%) for mortality and 0% (range 0% to 38%) for major complications. The percent of participants with >10% missing data was 1.0% for mortality and 1.3% for major complications. As a sensitivity analysis, we re-calculated participant-specific mortality and complication rates after

excluding records with missing data from the denominator. As shown in the figure below, there was high (>0.99) correlation between participant-specific rates calculated with missing data excluded versus imputed.



2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

These results suggest that our handling of missing outcome data is unlikely to impact performance results for the vast majority of participants.

2d. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

Note: *If empirical analyses do not provide adequate results—or are not conducted—justification must be provided and accepted in order to meet the must-pass criterion of Scientific Acceptability of Measure Properties. Each of the following questions has instructions if there is no empirical analysis.*

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To verify that each domain contributes statistical information but does not dominate the composite, we calculated the correlations between each domain-specific estimate and the overall composite score. These analyses were performed using data from July 2011 – June 2014.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; *if no empirical analysis, identify the components that were considered and the pros and cons of each*)

Pearson Correlation With Overall Composite

Mortality	Morbidity
0.74	0.89

The Pearson correlations were 0.74 for mortality versus overall composite measure and 0.89 for the morbidity domain score versus overall score.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components; if no empirical analysis, provide rationale for the components that were selected)*

Although risk-adjusted morbidity explains more of the variation in the overall composite score, it does not dominate. Both domains contribute statistical information.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used; if no empirical analysis, provide justification)*

To form the composite, we rescaled the morbidity and mortality domains by dividing by their respective standard deviations across STS participants and then added the two domains together. This weighting was then assessed by an expert panel to determine if it provided an appropriate reflection of the relative importance of the two domains. To facilitate the assessment, we calculated for a 1 percentage point change in mortality, what percentage point change in morbidity would be needed to achieve the same impact on the composite measure.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; if no empirical analysis, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

After rescaling, the relative weights in the final composite of risk-standardized mortality and risk-standardized major morbidity were 0.74 and 0.26, respectively. An implication of this weighting is that a 1 percentage point change in a participant's risk-adjusted mortality rate has the same impact as a 2.8 percentage point change in the site's risk-adjusted morbidity rate.

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting; if no empirical analysis, provide rationale for the selected rules for aggregation and weighting*)

This weighting was consistent with our expert panel's clinical assessment of each domain's relative importance.

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis,

depression score), Abstracted from a record by someone other than person obtaining original information (e.g., chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
Some data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
The STS Adult Cardiac Surgery Database (ACSD) has more than 1,100 participants, and local availability of data elements in electronic format will vary across institutions. Some institutions may have full EHR capability while others may have partial, or no availability. However, all data elements from participating institutions are submitted to the STS ACSD in electronic format following a standard set of data specifications. The majority of participating institutions obtain data entry software products that are certified for the purposes of collecting STS ACSD data elements.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
No feasibility assessment **Attachment:**

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
The data elements included in this measure have been standard in the STS Adult Cardiac Surgery Database for at least 3 years and some of them have been part of the database for more than 20 years. The variables are considered to be data elements that are readily available and already collected as part of the process of providing care.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.
Data Collection:
There are no additional costs for data collection specific to this measure for those presently using and participating in the STS Adult Cardiac Surgery Database. Costs to develop and maintain the measure included volunteer cardiothoracic surgeon time, STS staff time, and Duke Clinical Research Institute statistician and project management time.

Other fees:
STS Adult Cardiac Surgery Database participants (single cardiothoracic surgeons or a group of surgeons) pay annual participant fees of $3,500 if the majority of surgeons in the group are STS members and $4,750 if the majority of surgeons in the group are not STS members. In addition, there is a fee of $150 per member and $350 per nonmember for surgeons listed on the database Participation Agreement.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance

results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
Public Reporting Quality Improvement with Benchmarking (external benchmarking to multiple organizations) Quality Improvement (Internal to the specific organization)	

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

N/A

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

This is a new composite measure, which was developed in 2015 and will be published in 2016. STS plans to distribute participant-specific composite results in 2016 and roll out public reporting within the next year or so.

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

Please see 4a.2.

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

Data are provided in 1b.2 and 1b.4 as required.

4b.2. If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

N/A

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

We are not aware of any negative unintended consequences. All public reporting initiatives have the potential for unintended consequences, including gaming and risk aversion. We attempt to control the former through a careful audit process (in 2014, 10% of participants were audited) and the latter by having a robust methodology that appropriately adjusts the expected risk for providers who care for sicker patients.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0114 : Risk-Adjusted Postoperative Renal Failure
0115 : Risk-Adjusted Surgical Re-exploration
0119 : Risk-Adjusted Operative Mortality for CABG
0120 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR)
0121 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement
0122 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Replacement + CABG Surgery
0123 : Risk-Adjusted Operative Mortality for Aortic Valve Replacement (AVR) + CABG Surgery
0129 : Risk-Adjusted Postoperative Prolonged Intubation (Ventilation)
0130 : Risk-Adjusted Deep Sternal Wound Infection
0131 : Risk-Adjusted Stroke/Cerebrovascular Accident
1501 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair
1502 : Risk-Adjusted Operative Mortality for Mitral Valve (MV) Repair + CABG Surgery

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;
OR
The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):**
Are the measure specifications completely harmonized?

Yes

5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

N/A

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
N/A

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** STS_MVRR_Composite_Score_Appendix_-_S.4-S.11-S.14-S.15-_1b.2-_1b.4-_manuscripts.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** The Society of Thoracic Surgeons
Co.2 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-
Co.3 **Measure Developer if different from Measure Steward:** The Society of Thoracic Surgeons
Co.4 **Point of Contact:** Jane, Han, jhan@sts.org, 312-202-5856-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

- David Shahian, MD – Chair, Quality Measurement Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Gaetano Paone, MD – Chair, Task Force on Quality Initiatives; surgeon leader/clinical expert in adult cardiac surgery
- Richard S. D´Agostino, MD– Chair, Adult Cardiac Surgery Database Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Vinay Badhwar, MD – Chair, Public Reporting Task Force; surgeon leader/clinical expert in adult cardiac surgery
- Anthony P. Furnary, MD – Surgeon leader/clinical expert in adult cardiac surgery
- J. Scott Rankin, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Joseph C. Cleveland, Jr, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Jeffrey Jacobs, MD – Surgeon leader/clinical expert in congenital heart surgery
- Kristopher M George, MD – Surgeon leader/clinical expert in adult cardiac surgery
- Max He, MS – Statistician
- Sean O´Brien, PhD – Statistician
- Maria Grau-Sepulveda, MD – Statistician
- Jane Han, MSW – Staff, Senior Manager of Quality Metrics & Initiatives
- Donna McDonald, MPH, RN – Staff, STS Director of Quality

Members of the STS Task Force on Quality Initiatives and the Adult Cardiac Surgery Database Task Force provide clinical expertise as needed. The STS Workforce on National Database meets at the STS Annual Meeting and reviews measures on an annual basis.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2016
Ad.3 **Month and Year of most recent revision:** 06, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annually
Ad.5 **When is the next scheduled review/update for this measure?** 01, 2017

Ad.6 **Copyright statement:** N/A

Ad.7 **Disclaimers:** N/A
Ad.8 **Additional Information/Comments:** N/A

MEASURE WORKSHEET

This document summarizes the evaluation of the measure as it progresses through NQF's Consensus Development Process (CDP). The information submitted by measure developers/stewards is included after the Brief Measure Information, Preliminary Analysis, and Pre-meeting Public and Member Comments sections.

To navigate the links in the worksheet: Ctrl + click link to go to the link; ALT + LEFT ARROW to return

Brief Measure Information

NQF #: 0697
Measure Title: Risk Adjusted Case Mix Adjusted Elderly Surgery Outcomes Measure
Measure Steward: American College of Surgeons
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.
Developer Rationale: Reduced mortality and major morbidity rates for elderly following surgeries

Numerator Statement: 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.
Denominator Statement: 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)
Denominator Exclusions: Cases must first have ACS NSQIP eligible CPT codes on the submitted list of ~2900 codes. Major/multisystem trauma and transplant surgeries are excluded. Patients who are ASA 6 (brain-death organ donor) are not eligible surgical cases. Surgeries following within 30 d of an index procedure are an outcome (return to OR) and are not eligible to be new index cases. Thus, a patient known to have had a prior surgical operation within 30 days is excluded from having the subsequent surgery considered an index case.

Measure Type: Outcome
Data Source: Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Pharmacy, Electronic Clinical Data : Registry, Management Data, Paper Medical Records
Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 17, 2011 **Most Recent Endorsement Date:** Jan 17, 2011

Maintenance of Endorsement -- Preliminary Analysis

To maintain NQF endorsement endorsed measures are evaluated periodically to ensure that the measures still meets the NQF endorsement criteria ("maintenance"). The emphasis for maintaining endorsement is focused on how effective the measure is for promoting improvements in quality. Endorsed measures should have some experience from the field to inform the evaluation. The emphasis for maintaining endorsement is noted for each criterion.

Criteria 1: Importance to Measure and Report

1a. Evidence
Maintenance measures – less emphasis on evidence unless there is new information or change in evidence since the prior evaluation.

1a. Evidence. The evidence requirements for a health outcomes measure include providing rationale that supports the relationship of the health outcome to processes or structures of care. The guidance for evaluating the clinical evidence asks if the relationship between the measured health outcome and at least one clinical action is identified and supported by the stated rationale.

Summary of evidence:

- This maintenance measure, last reviewed in 2011, measures hospital-specific risk-adjusted mortality, a return to the operating room, or any of the morbidities as defined by American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP).
- The rationale provided by the developer to support this outcome measure addresses increased demand for surgical services among the aging population and the increased risk of post-operative morbidity and mortality within this group.
- The developer notes that in 2010 there were 51.4 million procedures performed in the US of which 19.2 million were performed on patients ages 65 and older and posits that a growing elderly population with increasing numbers of elderly persons undergoing surgical procedures, meaningful outcome measures are needed to mitigate the morbidity and mortality experienced by this high-risk population.
- The developer states that there are no national clinical practice guidelines specific for elderly patients undergoing surgical procedures. Instead, currently available national guidelines focus on medical conditions, general processes of care for hospitalized elderly, or occasionally, care related to a specific procedure.

Question for the Committee:

- *Does the Committee agree the underlying rationale for the measure remains reasonable?*
- *Does the Committee agree that the evidence supports the measure?*
- *Is there at least one thing that the provider can do to achieve a change in the measure results?*

Guidance from the Evidence Algorithm : Health outcome (Box 1)→relationship between outcome and at least one healthcare action identified/supported by stated rationale (Box 2) →Pass

Preliminary rating for evidence: ☒ **Pass** ☐ **No Pass**

1b. Gap in Care/Opportunity for Improvement and 1b. Disparities
Maintenance measures – increased emphasis on gap and variation

1b. Performance Gap. The performance gap requirements include demonstrating quality problems and opportunity for improvement.

The developer presented unpublished data obtained from an internal analysis of ACS NSQIP data:

- For 2014, there were 460 hospitals contributing 206,064 surgical cases on adults age 65 and over. ACS NSQIP reports results of application of the measure as an observed to expected (O/E) ratio. If the interval represented by the O/E ratio is entirely above 1.0, performance is identified as significantly worse than expected. If the interval is entirely below 1.0, performance is identified as significantly better than expected. Values that overlap 1 are interpreted as performing "as expected".
- The developer reports that for 2014, 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 ratio is reported as 0.55 to 1.6 in citing improvement.)
- 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. (Interquartile range differs slightly as reported in improvement.)
- The cited source identifies O/E ratio slopes for all hospitals included in its study as follows: improvement in mortality, 62% and morbidity, 70%.

Disparities:

The developer provided a summary of literature that addresses disparities in surgical care in older adults. These findings include:

- 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
- 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

- Individuals 80 years or older are 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
- Rate of surgical treatment of osteoarthritis of the knee in Medicare beneficiaries varies substantially by region of the country, sex, and race or ethnicity
- Racial disparities in the performance of coronary artery bypass grafts, carotid endarterectomy, and total hip replacement among Medicare beneficiaries persists

Questions for the Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *Are you aware of evidence of work done within the last 5 years that other disparities exist in this area of healthcare?*

Preliminary rating for opportunity for improvement: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a.

- Why do this just for older patients? Why not all?
- Why not keep PE in?
- Is composite score driven by the most common event (UTI?)?
- 49 high and 34 low outliers out of 460 hospitals (seems about right)
- Hospital level O/E reporting
- The rationale for the measure remains. There are several issues which will be listed in response to different comment sections. The evidence for the measure has not really changed. Although this intermediate outcome is a marker for quality - the need for increasingly sophisticated measures is increasing.
- The measure maybe too broad by lumping everyone over 65 together. That is not a physiological cutoff - reflects Medicare eligibility. Also, limited by including ASA 5.
- This is an OUTCOME measure and is a risk-adjusted OUTCOME of composite morbidity of adults over the age of 65 undergoing surgery.

1b.

- There has been much work done on social determinants (including NQF). These often apply to a younger population. The geriatric population has a group of different risk factors. A spouse of an 85 year old who is supposed to help in home care (in the 30 day window) maybe 88 and suffer from dementia- yet most social determinant scoring systems will state the patient lives with a care giver.
- The developer reports that there are both high outliers and low outliers for this measure demonstrating a continued performance gap.
- This measure can't be consistently implemented since it requires data abstraction and entry - requiring staff. This is one of the reasons that NSQIP is not utilized everywhere. There is no evidence that claims data alone or eCQM's will rpoduce validated results.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

2a1. Reliability Specifications

Maintenance measures – no change in emphasis – specifications should be evaluated the same as with new measures

2a1. Specifications requires the measure, as specified, to produce consistent (reliable) and credible (valid) results about the quality of care when implemented.

Data source(s):

- Electronic Clinical Data, Electronic Health Record, Imaging/Diagnostic Study, Laboratory, Pharmacy Registry, Management Data, Paper Medical Records

Specifications:

- The measure is specified as a facility-level measure for ambulatory surgery center and hospital/acute care

facility settings. Better quality – lower score.

- The denominator includes patients who are 65 years of age or older who undergo any of the ACS NSQIP listed (CPT) surgical procedures (about 2,900). An appendix of ACS NSQIP code inclusions is provided.
- The numerator includes: hospital-specific risk-adjusted mortality; return to the operating room; or any of the following morbidities:
  - Cardiac Arrest requiring CPR,
  - Myocardial Infarction,
  - Sepsis, Septic Shock,
  - Deep Incisional Surgical Site Infection,
  - 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.
- This outcome measure is risk adjusted, using a statistical risk model; though, it is unclear which risk factors are actually included in the measure that is being put forward for endorsement.

Questions for the Committee :

- *Are all the data elements clearly defined? Are all appropriate codes included?*
- *Absent a calculation algorithm, is the logic clear and compelling?*
- *Is it likely this measure can be consistently implemented?*

2a2. Reliability Testing Testing attachment
Maintenance measures – less emphasis if no new testing data provided

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise enough to distinguish differences in performance across providers.

For maintenance measures, summarize the reliability testing from the prior review:

- Previous reliability testing used intra-class correlation coefficient (ICC) for hospitals.
  - Using a minimum acceptable reliability for mortality/serious morbidity in patients >= 65 of 0.4 (intra-class correlation interpreted as "moderate") for which about 180 cases are required, 80.8% of US hospitals and 84.8% of ACS NSQIP met the 0.4 requirement at the score level.

Describe any updates to testing

- Reliability of the measure score was assessed using what is described as a standard method/approach and equations derived from the 2011 – 2014 data set applied to hospital data for one year (2014, 460 hospitals, 206,064 cases).

SUMMARY OF TESTING

Reliability testing level ☒ **Measure score** ☐ **Data element** ☐ **Both**

Reliability testing performed with the data source and level of analysis indicated for this measure ☒ **Yes** ☐ **No**

Method(s) of reliability testing

- Reliability was assessed using information from a random intercept, fixed slope, hierarchical logistic regression model and applying the Spearman-Brown prophecy formula.

Results of reliability testing

- Reliability was evaluated for 460 hospitals collecting data during 2014. The mean number of cases per hospital in the 2014 data set was 448, but there was positive skew in the distribution of sample sizes (median=437).
- The developer 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).

- Based on a regression analysis, the developer determined that an empirical estimate of the sample size required to achieve reliability of 0.4 is 115. The developer notes that 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."

The reliability of the 2014 dataset was examined under 4 conditions defined by the combination of VTE (included or not included) and SES variables (included or not included). Over 80% of hospitals had a "minimally acceptable" reliability estimate for the measure across all 4 variations. 83.92% of hospitals met the "minimally acceptable" estimate for the model recommended for the measure (See below)

Calibration range	Percent without VTE, without SES
0.00-0.20	10.43
0.21-0.40	5.65
0.41-0.60	14.13
0.61-0.80	46.96
0.81-1.00	22.83

Questions for the Committee:

- *Is the test sample adequate to generalize for widespread implementation?*
- *Is the method of testing and interpretation of results clear and compelling?*
- *Do the results demonstrate sufficient reliability so that differences in performance can be identified?*

Guidance from the Reliability Algorithm

Precise specifications (Box 1) → empiric reliability testing (Box 2) → Testing at measure score (Box 4) → Method described (Box 5) → Confidence in measure score →moderate

Preliminary rating for reliability: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

2b. Validity
Maintenance measures – less emphasis if no new testing data provided

2b1. Validity: Specifications

2b1. Validity Specifications. This section should determine if the measure specifications are consistent with the evidence.

Specifications consistent with evidence in 1a. ☐ **Yes** ☒ **Somewhat** ☐ **No**

Question for the Committee:

- *Are the specifications consistent with the evidence?*

2b2. Validity testing

2b2. Validity Testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

For maintenance measures, summarize the validity testing from the prior review:

In the previous review, model validity (c-statistic, discrimination) using ACS NSQIP Data from 2008 and applied to 2007 data was provided.

Describe any updates to validity testing
The developer describes the risk model diagnostics approach to evaluation of risk model quality and processes. While important information, the demonstration of risk model adequacy does not meet NQF requirements for validity testing of the measure score.

SUMMARY OF TESTING
Validity testing level ☐ **Measure score** ☐ **Data element testing against a gold standard** ☐ **Both**

Method of validity testing of the measure score:
- ☐ **Face validity only**
- ☐ **Empirical validity testing of the measure score**

Validity testing method: N/A

Validity testing results: N/A

Questions for the Committee:
- *What approach would the Committee accept to receive and evaluate validity testing in the current cycle?*

2b3-2b7. Threats to Validity

2b3. Exclusions:
The developer notes patients are excluded from the measure for the following reasons:
- Surgeries are not on the ACS NSQIP CPT code list;
- Major/multisystem trauma and transplant surgeries
- ASA 6 (brain-death organ donor)
- Prior surgical procedures within 30 days of a potential index procedure

Information regarding number of excluded cases is not provided thus cannot appreciate analysis of exclusions.

Questions for the Committee:
- *Are the exclusions consistent with the evidence?*
- *Are any patients or patient groups inappropriately excluded from the measure?*
- *Are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)?*

2b4. Risk adjustment: **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

Conceptual rationale for SDS factors included ? ☐ **Yes** ☒ **No**

SDS factors included in risk model? ☐ **Yes** ☒ **No**

Risk adjustment summary
The developer reports:
- Statistical risk modeling is performed in a step-wise fashion – case mix adjustment, variable selection, then risk adjustment.
- For variable selection of risk factors, logistic regression is performed using NSQIP predictors demonstrating statistical significance (P<0.05) from which a subset is chosen to create a predictor set. The original set of predictors is included. For this measure the 3 predictors are ASA class, CPT risk and functional status .
- Additional variables modeled to explore SDS were median income, Hispanic ethnicity, race. Inclusion/exclusion of VTE was also assessed.
  - The developer noted that addition of SDS factors are not influential in risk adjustment (weighted kappa = 0.9287) and that removal of VTE has important effect on outlier and decile status (weighted kappa = 0.5982)
- The developer reports that C-statistic has been used to evaluate discrimination of the risk model and is applied to evaluation of exclusion/inclusion of VTE and SDS factors in the current measure. Other tests to demonstrate accuracy of prediction about probability (Brier), and goodness of fit for models assessing observed and expected

rates (Hosmer-Lemeshow), were computed for the 2011-2014 dataset
- The developer reports that model quality remains consistent when the 2011-2014 equations are applied to the 2010 dataset.

Questions for the Committee:
- *Is the risk-adjustment strategy included in the measure clear, complete and appropriate for the measure?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*
- *Are all of the risk adjustment variables present at the start of care? If not, describe the rationale provided.*
- *Is there a conceptual rationale to include (or not include) SDS factors in the risk adjustment approach? If so, are the relevant risk factors considered?*
- *Do you agree with the developer's decision not to include SDS factors in the risk-adjustment approach?*

2b5. Meaningful difference *(can statistically significant and clinically/practically meaningful differences in performance measure scores can be identified)*:

The default methodology for discrimination performance is based on the computed 95% CI (using Ulm's method) for the O/E ratio. If the interval is entirely above1.0, the hospital is identified as having performance significantly worse than expected. If the interval is entirely below 1.0, the hospital is identified as having performance significantly better than expected. If the interval overlaps 1.0 the hospital is performing "as expected."

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.

Question for the Committee:
- *Does this measure identify meaningful differences about quality?*

2b6. Comparability of data sources/methods:

Participating hospitals collect and report data to ACS NSQIP. Hospitals that do not participate can submit data to the implementing organization.

Question for the Committee:
Will data submitted by non-participating hospitals be comparable, yielding results comparable to those reported to NSQIP?

2b7. Missing Data

The developer reports high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Guidance from the Validity Algorithm:
Specifications consistent with the evidence (Box 1) → Potential threats to validity addressed (Box 2)

Preliminary rating for validity: ☐ **High** ☐ **Moderate** ☐ **Low** ☒ **Insufficient**
Preliminary rating is based on lack of clarity about risk adjustment strategy, testing, and comparability of data sources/methods.

Committee pre-evaluation comments

Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2d)

2a.
- Based on previous assessments agree that there is adequate reliability.
- This measure is limited by the limits of the ACS NSQIP:
  1)Not all hospitals participate

2)Participating hospitals submit a sample of cases, not all cases performed
3)The risk-adjustment does not have measures of preoperative cognition or frailty that would impact postoperative outcomes in older adults
4)Postoperative delirium is not measured in the ACS NSQIP
5)The composite outcome includes a wide range of very morbid things (unplanned reoperation, cardiac arrest, sepsis, septic shock, deep incisional SSI, organ space SSI, unplanned re-intubation, pneumonia, renal failure) but also includes postoperative UTI. The high occurrence of postoperative UTI with low morbidity could drive this measure more than the more morbid outcomes. Additionally, this measure excludes postoperative superficial SSI

2a2.

- Yes
- The reliability of the data elements entered into the ACS NSQIP has been published and is reliable.
- The developers also conducted reliability testing at the score level.
- A sample size of 180 cases per hospital annual is estimated for hospitals to have a 0.4 (moderate level) of reliability. 90.8% of hospitals and 84.8% of ACS NSQIP hospitals are able to achieve this volume.
- 2b1.
- c-statistic for the models from 2007 were the basis for original measure endorsement.
- The developers investigated the role of excluding VTE and including or excluding SES in their risk adjustment models and include this data on page 26 of the measure application.
- The developers conclude that the model was not impacted by including SES.

2b2.

- Although valid at the NSQIP sites - need to test at non NSQIP sites using claims data. That data will require primary validity checking.
- Developers conducted extensive risk model diagnostic testing (described in published article "A Comprehnsive Evaluation of Statistical Reliability in ACS NSQIP profiling Models".

2b3.

- The exclusions are consistent with the evidence - but maybe too broad. None are inappropriately excluded - but may need to exclude more. The results support the risk adjustment model.
- Need to include social determinants as mentioned above.
- The risk adjustment methodology depends on strict definitions for the potential outcomes. This works for a NSQIP hospital - but probably won't where administrative data is utilized without validity checking. Coding is too inaccurate as often financially driven
- Does produce meaningful data for participating hospitals"
- This measure is limited by the limits of the ACS NSQIP:

1) Not all hospitals participate
2) Participating hospitals submit a sample of cases, not all cases performed
3) The risk-adjustment does not have measures of preoperative cognition or frailty that would impact postoperative outcomes in older adults
4) Postoperative delirium is not measured in the ACS NSQIP
5) The composite outcome includes a wide range of very morbid things (unplanned reoperation, cardiac arrest, sepsis, septic shock, deep incisional SSI, organ space SSI, unplanned re-intubation, pneumonia, renal failure) but also includes postoperative UTI. The high occurrence of postoperative UTI with low morbidity could drive this measure more than the more morbid outcomes. Additionally, this measure excludes postoperative superficial SSI.

Criterion 3. Feasibility
Maintenance measures – no change in emphasis – implementation issues may be more prominent

3. Feasibility is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

The developer noted the following:

- Data is generated as byproduct of care processes during care delivery.
- Coding/abstraction performed by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or entry into registry).
  - Historically, trained data collectors within ACS NSQIP and a comprehensive support system have facilitated data entry into NSQIP.
- No data elements are in defined fields in electronic sources.
- A completely electronic medical record (EMR) would be needed to capture all risk factors that enter into the

model. In addition, a software module (currently available to ACS NSQIP subscribers) will be required to transfer information from the EMR to a measure submission database. The ACS NSQIP is in the process of developing an automated process with EMR vendors, however, electronic entry for this measure is not currently available.
- ACS NSQIP has been open to subscription by private sector hospitals since 2004 and over 600 hospitals participate.
- Program participants are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Questions for the Committee:

- *Are the required data elements routinely generated and used during care delivery?*
- *Is the data collection strategy ready to be put into operational use within and outside the ACS NSQIP?*

Preliminary rating for feasibility: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 3: Feasibility

- Yes - but require chart abstraction. Not all EHR's capture this data.
- Not ready for non-NSQIP hospitals
- While participation in ACS NSQIP does present a burden at the data entry level, I would argue that hospitals invested in ACS NSQIP participation do benefit from the program.

Criterion 4: Usability and Use

Maintenance measures – increased emphasis – much greater focus on measure use and usefulness, including both impact /improvement and unintended consequences

4. Usability and Use evaluate the extent to which audiences (e.g., consumers, purchasers, providers, policymakers) use or could use performance results for both accountability and performance improvement activities.

Current uses of the measure
Publicly reported? ☒ **Yes** ☐ **No**

Current use in an accountability program? ☒ **Yes** ☐ **No**

Accountability program details :

Public Reporting: Hospital Compare – 131 hospitals currently report their risk-adjusted surgery outcomes data for NQF-endorsed measures from ACS.

Quality Improvement – internal, and external with benchmarking: ACS NSQIP, a voluntary program with 600 participating hospitals, provides hospitals with clinical data specific to their organization with which to track and improve outcomes internally, as well as benchmarked, risk-adjusted outcomes reports with which to compare themselves to other institutions.

Improvement results:

The developer reports a recent analysis indicates that over 8 years in the program, 62% and 71% of hospitals improve their performance in mortality and risk-adjusted complications. Annual reductions are approximately 0.8% in mortality and 3.1% in morbidity.

For 2014, there were 460 hospitals contributing 206,064 surgeries on adults age 65 and older.

Unexpected findings (positive or negative) during implementation:
The developer does not report any unexpected findings

Potential harms:
The developer does not report any potential harms

Feedback: At prior review, Committee members highlighted data abstraction burden and need to conform to NSQIP methodology as challenges to feasibility for non-NSQIP hospitals; however, the Committee expressed belief that burden was offset by having a cross-cutting measure on outcomes.

***Questions for the Committee**:*

- *How can the performance results be used to further the goal of high-quality, efficient healthcare?*
- *Do the benefits of the measure outweigh any potential unintended consequences?*

Preliminary rating for usability and use: ☐ **High** ☒ **Moderate** ☐ **Low** ☐ **Insufficient**

Committee pre-evaluation comments

Criteria 4: Usability and Use

- Not clear that any hospital has used this measure to drive improvement. It needs to be sliced and diced to subgroups to be meaningful. It is a marker as an intermediate outcome, but with the many confounding issues as above and the extreme variability of the patient population, this represents a measure that will be misinterpreted if publically reported.
- 131 ACS NSQIP hospitals currently report data risk-adjusted data.

Criterion 5: Related and Competing Measures

Related or competing measures

0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB). Developer: ACS

0706 : Risk Adjusted Colon Surgery Outcome Measure. Developer: ACS

Harmonization

The developer reports that the measure specifications are completely harmonized with the related measures.

Pre-meeting public and member comments

-

NATIONAL QUALITY FORUM

NQF #: 0697 **NQF Project:** Patient Outcomes Measures: Phases I and II

1. IMPACT, OPPORTUITY, EVIDENCE - IMPORTANCE TO MEASURE AND REPORT

Importance to Measure and Report is a threshold criterion that must be met in order to recommend a measure for endorsement. All three subcriteria must be met to pass this criterion. See guidance on evidence.

Measures must be judged to be important to measure and report in order to be evaluated against the remaining criteria. (evaluation criteria)

1c.1 Structure-Process-Outcome Relationship *(Briefly state the measure focus, e.g., health outcome, intermediate clinical outcome, process, structure; then identify the appropriate links, e.g., structure-process-health outcome; process- health outcome; intermediate clinical outcome-health outcome)*:

This is a risk-adjusted outcome measure. The population is rapidly aging, with the proportion of adults aged 65 and older projected to grow from 15% in 2015 to 24% of the population in 2060. (U.S. Census Bureau 2014) The aging of the population will significantly increase the demand for surgical services. In 2010 there were 51.4 million procedures performed across the United States, with 19.2 million of those performed on patients ages 65 and older.(National Hospital Discharge Survey 2010) Given the growth of the population aged 65 years and older, and the substantial risk for post-operative morbidity and mortality within this group, additional resources should be directed toward assessing and improving outcomes among elderly surgical patients.

Due to the disproportionate need for surgical procedures in the elderly, the surgical workload is forecast to increase in general surgery, orthopedics, urology, and neurosurgery by 31%, 28%, 35%, and 28%, respectively.(Etzione et al. 2003) Furthermore, oncologic surgery among patients aged 80 years and older is anticipated to increase by 51%. (Etzione et al. 2003) Similarly, overall rates of general surgery procedures for individuals aged 65 years and older have been estimated to be three times higher than those aged 15 to 44 years and 1.6 times higher than rates for those aged 45 to 64 years. Colon resection is the most rapidly increasing operation out of five common general surgery procedures (laparoscopic cholecystectomy, appendectomy, inguinal hernia, and breast excision); in individuals 65 years and older colon resection has increased at a rate that is 17-fold higher than for those aged 15 to 44 years and a 4-fold higher rate than in those aged 45 to 64 years.(Liu et al. 2004) Between 1994 and 2003, total discharges after lung, esophageal, and pancreatic surgery in patients aged 80 years and older increased by 76%.(Finlayson et al. 2007) With a growing portion of the population becoming elderly, and increasing numbers of elderly persons undergoing surgical procedures, meaningful outcome measures are needed to mitigate the morbidity and mortality experienced by this high-risk population.

Older adults experience significantly higher rates of postoperative morbidity and mortality. (Polanczyk et al. 2001; Finlayson, Birkmeyer. 2001; Hamel et al. 2005; Turrentine et al. 2006; Bentrem et al. 2009; Sheetz et al. 2014) Postoperative morbidity includes 1.2-2 times increased risks for cardiac, pulmonary and urologic events as well as 2.9-6.7 times the risk for death.(Bentrem et al. 2009) Furthermore, older adults experience a decreased ability to recover once complications occur, making it important to identify and recognize postoperative complications. (Sheetz et al. 2014) Though processes of care for elderly surgical patients have been evaluated through the RAND/UCLA Appropriateness Methodology (see 1c.4 below), the level of evidence remains limited and the value of assessing outcomes cannot be underestimated.

Risk-adjusted postoperative morbidity and mortality rates at the hospital-level remain a critical piece of information in order to assess hospital quality and promote improvement in this vulnerable population.

1c.2-3 Type of Evidence *(Check all that apply)*:

Observational based on prospectively collected rigorously controlled data (ACS NSQIP)

1c.4 Directness of Evidence to the Specified Measure *(State the central topic, population, and outcomes addressed in the body of evidence and identify any differences from the measure focus and measure target population)*:

Despite evidence that the elderly require specific perioperative care, (Pofahl et al. 2003; Rosenthal et al. 2004; Loran et a. 2005; Jin, Chung. 2001) most previous initiatives to delineate process-based quality indicators for the elderly did not focus on surgical populations.(Shekelle et al. 2001) In recent years, increased attention has turned to elderly patients undergoing surgical procedures, with significant contributions from members of the Research ANd Development (RAND) Corporation; their research has identified process measures that are specific to the elderly undergoing surgery and include but are not limited to screens for nutrition, cognition, and delirium risk.(McGory et al. 2005; McGory et al. 2009) Using the RAND/University of California Los Angeles Appropriateness Methodology, a modified Delphi technique, quality indicators were developed for elderly patients undergoing elective major abdominal surgery (McGory et al. 2005) and for elderly patients undergoing ambulatory or inpatient surgery for nearly all surgical specialties that care for elderly patients (cardiothoracic, colorectal, general, gynecology, orthopedic, urology, and vascular surgery).(McGory et al. 2009) The need for additional research for surgical quality improvement that focuses on elderly care is highlighted by the fact that most of the quality indicators identified by McGory et al. did not have level one randomized controlled trial evidence. Indeed, the RAND/UCLA Appropriateness Methodology was designed to identify as accurately as possible the highest quality processes when the highest level of evidence is not available.(Brook. 1994)

In sum, there is a lack of high level evidence for process measures in elderly surgery, supporting the need to maintain outcome-based measures to evaluate quality of surgical care in the elderly.

1c.5 Quantity of Studies in the Body of Evidence *(Total number of studies, not articles)*:

1c.6 Quality of Body of Evidence (*Summarize the certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence resulting from study factors. Please address: a) study design/flaws; b) directness/indirectness of the evidence to this measure (e.g., interventions, comparisons, outcomes assessed, population included in the evidence); and c) imprecision/wide confidence intervals due to few patients or events):*

1c.7 Consistency of Results across Studies *(Summarize the consistency of the magnitude and direction of the effect):*

1c.8 Net Benefit *(Provide estimates of effect for benefit/outcome; identify harms addressed and estimates of effect; and net benefit - benefit over harms)*:

1c.9 Grading of Strength/Quality of the Body of Evidence. Has the body of evidence been graded?

1c.10 If body of evidence graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.11 System Used for Grading the Body of Evidence: N/A

1c.12 If other, identify and describe the grading scale with definitions:

1c.13 Grade Assigned to the Body of Evidence: Not applicable to outcomes based measures.

1c.14 Summary of Controversy/Contradictory Evidence: There are no specific process measures that have Level I evidence for the evaluation of surgery in the elderly. There are general process measures (i.e. SCIP) that are applicable to surgery in the elderly, however, internal analyses using clinical risk adjusted ACS NSQIP data demonstrate little to no correlation of these processes to outcomes. This dilemma/circumstance supports the need to use outcomes based measures.

1c.15 Citations for Evidence other than Guidelines*(Guidelines addressed below)*:

Anderson R, DeTurk P. United States Life Tables, 1999. Natl Vital Stat Rep. 2002;50(6):33.

Bentrem, D. J., M. E. Cohen, D. M. Hynes, C. Y. Ko and K. Y. Bilimoria (2009). "Identification of specific quality improvement opportunities for the elderly undergoing gastrointestinal surgery.

Brook RH, Chassin MR, Fink A, Solomon DH, Kosecoff J, Park RE. A method for the detailed assessment of the appropriateness of medical technologies. Int J Technol Assess Health Care. 1986;2(1):53-63.

Brook RH. The RAND/UCLA Appropriateness Method. Clinical practice guideline development: methodology perspectives. Rockville, MD: Pulblic Helath Service: ACHR; 1994.

Etzioni DA, Liu JH, Maggard MA, Ko CY. The aging population and its impact on the surgery workforce. Ann Surg. Aug 2003;238(2):170-177.

Etzioni DA, Liu JH, O'Connell JB, Maggard MA, Ko CY. Elderly patients in surgical workloads: a population-based analysis. Am Surg. Nov 2003;69(11):961-965.

Finlayson E, Fan Z, Birkmeyer JD. Outcomes in octogenarians undergoing high-risk cancer operation: a national study. J Am Coll Surg. Dec 2007;205(6):729-734.

Finlayson EV, Birkmeyer JD. Operative mortality with elective surgery in older adults. Eff Clin Pract. Jul-Aug 2001;4(4):172-177.

Hamel MB, Henderson WG, Khuri SF, Daley J. Surgical outcomes for patients aged 80 and older: morbidity and mortality from major noncardiac surgery. J Am Geriatr Soc. Mar 2005;53(3):424-429.

Jin F, Chung F. Minimizing perioperative adverse events in the elderly. Br J Anaesth. Oct 2001;87(4):608-624.

Liu JH, Etzioni DA, O'Connell JB, Maggard MA, Ko CY. The increasing workload of general surgery. Arch Surg. Apr 2004;139(4):423-428.

Loran DB, Hyde BR, Zwischenberger JB. Perioperative management of special populations: the geriatric patient. Surg Clin North Am. Dec 2005;85(6):1259-1266, xi.

McGory ML, Kao KK, Shekelle PG, et al. Developing quality indicators for elderly surgical patients. Ann Surg. Aug 2009;250(2):338-347.

McGory ML, Shekelle PG, Rubenstein LZ, Fink A, Ko CY. Developing quality indicators for elderly patients undergoing abdominal

operations. J Am Coll Surg. Dec 2005;201(6):870-883.
Pofahl WE, Pories WJ. Current status and future directions of geriatric general surgery. J Am Geriatr Soc. Jul 2003;51(7 Suppl):S351-354.
Polanczyk CA, Marcantonio E, Goldman L, et al. Impact of age on perioperative complications and length of stay in patients undergoing noncardiac surgery. Ann Intern Med. Apr 17 2001;134(8):637-643.
Rosenthal RA, Kavic SM. Assessment and management of the geriatric patient. Crit Care Med. Apr 2004;32(4 Suppl):S92-105.
Sheetz, K. H., R. W. Krell, M. J. Englesbe, J. D. Birkmeyer, D. A. Campbell, Jr. and A. A. Ghaferi (2014). "The importance of the first complication: understanding failure to rescue after emergent surgery in the elderly." J Am Coll Surg 219(3): 365-370.
Shekelle PG, MacLean CH, Morton SC, Wenger NS. Assessing care of vulnerable elders: methods for developing quality indicators. Ann Intern Med. Oct 16 2001;135(8 Pt 2):647-652.
Turrentine FE, Wang H, Simpson VB, Jones RS. Surgical risk factors, morbidity, and mortality in elderly patients. J Am Coll Surg. Dec 2006;203(6):865-877.

1c.16 Quote verbatim, the specific guideline recommendation *(Including guideline # and/or page #)*:
While quality indicators for patients aged 65 years or older undergoing surgical procedures exist (as described above), there are no national clinical practice guidelines specific for elderly patients undergoing surgical procedures. Instead, currently available national guidelines focus on medical conditions (e.g., management of chronic heart failure), general processes of care for hospitalized elderly (e.g., prevention of pressure ulcers), or occasionally care related to a specific procedure (e.g., hip fracture). (http://www.guideline.gov/; accessed 5/16/2016).

1c.17 Clinical Practice Guideline Citation: N/A

1c.18 National Guideline Clearinghouse or other URL: N/A

1c.19 Grading of Strength of Guideline Recommendation. Has the recommendation been graded?

1c.20 If guideline recommendation graded, identify the entity that graded the evidence including balance of representation and any disclosures regarding bias:

1c.21 System Used for Grading the Strength of Guideline Recommendation: N/A

1c.22 If other, identify and describe the grading scale with definitions:

1c.23 Grade Assigned to the Recommendation: N/A

1c.24 Rationale for Using this Guideline Over Others: N/A

Based on the NQF descriptions for rating the evidence, what was the developer's assessment of the quantity, quality, and consistency of the body of evidence?
1c.25 Quantity: **1c.26** Quality: **1c.27** Consistency**:**

1. Evidence, Performance Gap, Priority – Importance to Measure and Report

Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance. ***Measures must be judged to meet all subcriteria to pass this criterion and be evaluated against the remaining criteria.***

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form
0697_Evidence_Maintenance-May2016.doc

1b. Performance Gap
Demonstration of quality problems and opportunity for improvement, i.e., data demonstrating:
- considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or

- disparities in care across population groups.

1b.1. Briefly explain the rationale for this measure *(e.g., the benefits or improvements in quality envisioned by use of this measure)*
Reduced mortality and major morbidity rates for elderly following surgeries.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for endorsement maintenance. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included). This information also will be used to address the subcriterion on improvement (4b.1) under Usability and Use.*
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

1b.3. If no or limited performance data on the measure as specified is reported in 1b2, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement.
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.

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for endorsement maintenance. 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 subcriterion on improvement (4b.1) under Usability and Use.*
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)

1b.5. If no or limited data on disparities from the measure as specified is reported in 1b4, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations.
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.

1c. High Priority (previously referred to as High Impact)

The measure addresses:

- a specific national health goal/priority identified by DHHS or the National Priorities Partnership convened by NQF; OR
- a demonstrated high-priority (high-impact) aspect of healthcare (e.g., affects large numbers of patients and/or has a substantial impact for a smaller population; leading cause of morbidity/mortality; high resource use (current and/or future); severity of illness; and severity of patient/societal consequences of poor quality).

1c.1. Demonstrated high priority aspect of healthcare

Affects large numbers, Severity of illness, Frequently performed procedure, Leading cause of morbidity/mortality, Patient/societal consequences of poor quality, High resource use

1c.2. If Other:

1c.3. Provide epidemiologic or resource use data that demonstrates the measure addresses a high priority aspect of healthcare. List citations in 1c.4.

The population is rapidly aging, with the proportion of adults aged 65 and older projected to grow from 15% in 2015 to 24% of the population in 2060. (U.S. Census Bureau 2014) The aging of the population will significantly increase the demand for surgical services. In 2010 there were 51.4 million procedures performed across the United States, with 19.2 million of those performed on patients ages 65 and older.(National Hospital Discharge Survey 2010)

In addition to consuming more medical resources than their younger counterparts, the elderly are at greater risk of morbidity and mortality after surgical procedures. Major perioperative complications after non-emergency surgery have been documented at increasing rates with increasing age: 5.7% of patients 60-69 years of age, 9.6% of patients 70 to 79 years of age, and 12.5% of patients at least 80 years of age.(Polanczyk et al. 2001) Finlayson et al. demonstrated that the mortality associated with 14 major elective surgeries in Medicare beneficiaries ranged from 1.3% to 13.7% depending on the procedure, with the highest mortality associated with mitral valve replacement (10.5%), esophagectomy (13.6%), and pneumonectomy (13.7%). Those older than 80 years of age are at particularly high risk for perioperative adverse events.(Finlayson, Birkmeyer. 2001; Hamel et al. 2005; Turrentine et al. 2006) Finalyson et al. demonstrated that the operative mortality among octogenarians was significantly higher than that of their younger counterparts (patients aged 65 to 69 years) for esophagectomy (19.9% versus 8.8%, $p < 0.0001$), pancreatectomy (15.5% versus 6.7%, $p < 0.0001$), and lung resections (6.9% versus 3.7%, $p < 0.0001$) for cancer.(Finlayson et al. 2007)

Recent literature continues to demonstrate poor outcomes among aging individuals. Older adults experience 2.9 to 6.7 times higher rates of postoperative mortality compared to their younger counterparts after adjusting for comorbid conditions; similarly older adults experience 1.2 to 2 times higher rates of postoperative complications, including postoperative cardiac, pulmonary and urologic complications.(Bentrem, Cohen et al. 2009) With advancing age, patients are at increased risk for geriatric syndromes such as frailty, which is characterized by decreased reserve and ability to respond to physiologic stress such as surgery. Frailty among elderly individuals may contribute to this excess postoperative morbidity and mortality.(Anaya, Johanning et al. 2014, Robinson, Walston et al. 2015) Older individuals do poorly when complications do occur. Early recognition of postoperative complications is critical in order to “rescue” patients from subsequent decline.(Sheetz, Krell et al. 2014)

In addition to causing patients significant harm and potentially leading to death, postoperative adverse events are associated with a significant financial burden. The cost of ventilator associated pneumonia has been documented to be between $10,019 and $57,158 with the daily cost of intensive care unit care being $1,861. (Warren et al. 2003; Safdar et al. 2005; Cocanour et al. 2005) The cost of postoperative acute renal failure ranges from $18,414 to $25,219. (Pronovost et al. 2001; Dimick et al. 2003) Given the increasing cost of health care and the need to improve value in care delivery, reducing complications can avert significant costs associated with surgical care for elderly adults.

Individuals aged 65 years and older consume a significant proportion of health care resources and surgical procedures, resource use which is magnified when postoperative complications occur. The elderly carry increased risk for postoperative morbidity and mortality. Reductions in postoperative morbidity and mortality will not only improve patient well-being but will reduce the cost of medical care.

1c.4. **Citations for data demonstrating high priority provided in 1a.3**

Anaya, D. A., J. Johanning, S. A. Spector, M. R. Katlic, A. C. Perrino, J. Feinleib and R. A. Rosenthal (2014). "Summary of the panel session at the 38th Annual Surgical Symposium of the Association of VA Surgeons: what is the big deal about frailty?" JAMA Surg 149(11): 1191-1197.
Bentrem, D. J., M. E. Cohen, D. M. Hynes, C. Y. Ko and K. Y. Bilimoria (2009). "Identification of specific quality improvement opportunities for the elderly undergoing gastrointestinal surgery.
Centers for Disease Control, National Hospital Discharge Survey: Number of all-listed procedures for discharges from short-stay hospitals, by procedure category and age: United States, 2010, 2010, http://www.cdc.gov/nchs/data/nhds/4procedures/2010pro4_numberprocedureage.pdf.
Cocanour CS, Ostrosky-Zeichner L, Peninger M, et al. Cost of a ventilator-associated pneumonia in a shock trauma intensive care unit. Surg Infect (Larchmt). Spring 2005;6(1):65-72.
Dimick JB, Pronovost PJ, Cowan JA, Lipsett PA. Complications and costs after high-risk surgery: where should we focus quality improvement initiatives? J Am Coll Surg. May 2003;196(5):671-678.
Finlayson E, Fan Z, Birkmeyer JD. Outcomes in octogenarians undergoing high-risk cancer operation: a national study. J Am Coll Surg. Dec 2007;205(6):729-734.
Hamel MB, Henderson WG, Khuri SF, Daley J. Surgical outcomes for patients aged 80 and older: morbidity and mortality from major noncardiac surgery. J Am Geriatr Soc. Mar 2005;53(3):424-429.
Polanczyk CA, Marcantonio E, Goldman L, et al. Impact of age on perioperative complications and length of stay in patients undergoing noncardiac surgery. Ann Intern Med. Apr 17 2001;134(8):637-643.
Pronovost P, Garrett E, Dorman T, et al. Variations in complication rates and opportunities for improvement in quality of care for patients having abdominal aortic surgery. Langenbecks Arch Surg. Jul 2001;386(4):249-256.
Robinson, T. N., J. D. Walston, N. E. Brummel, S. Deiner, C. H. t. Brown, M. Kennedy and A. Hurria (2015). "Frailty for Surgeons: Review of a National Institute on Aging Conference on Frailty for Specialists." J Am Coll Surg 221(6): 1083-1092.
Safdar N, Dezfulian C, Collard HR, Saint S. Clinical and economic consequences of ventilator-associated pneumonia: a systematic review. Crit Care Med. Oct 2005;33(10):2184-2193.
Sheetz, K. H., R. W. Krell, M. J. Englesbe, J. D. Birkmeyer, D. A. Campbell, Jr. and A. A. Ghaferi (2014). "The importance of the first complication: understanding failure to rescue after emergent surgery in the elderly." J Am Coll Surg 219(3): 365-370.
Turrentine FE, Wang H, Simpson VB, Jones RS. Surgical risk factors, morbidity, and mortality in elderly patients. J Am Coll Surg. Dec 2006;203(6):865-877.
United States Census Bureau, National Population Projections: Table 6. Percent Distribution of the Projected Population by Sex and Selected Age Groups for the United States: 2015 to 2060, 2014, https://www.census.gov/population/projections/data/national/2014/summarytables.html
Warren DK, Shukla SJ, Olsen MA, et al. Outcome and attributable cost of ventilator-associated pneumonia among intensive care unit patients in a suburban medical center. Crit Care Med. May 2003;31(5):1312-1317.

1c.5. **If a PRO-PM (e.g. HRQoL/functional status, symptom/burden, experience with care, health-related behaviors), provide evidence that the target population values the measured PRO and finds it meaningful.** *(Describe how and from whom their input was obtained.)*

2. Reliability and Validity—Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. ***Measures must be judged to meet the subcriteria for both reliability and validity to pass this criterion and be evaluated against the remaining criteria.***

2a.1. Specifications The measure is well defined and precisely specified so it can be implemented consistently within and across organizations and allows for comparability. eMeasures should be specified in the Health Quality Measures Format (HQMF) and the Quality Data Model (QDM).

De.5. **Subject/Topic Area** *(check all the areas that apply):*
Surgery : General Surgery

De.6. **Cross Cutting Areas** *(check all the areas that apply):*

S.1. **Measure-specific Web Page** *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*
NA

S.2a. **If this is an eMeasure**, HQMF specifications must be attached. Attach the zipped output from the eMeasure authoring tool (MAT) - if the MAT was not used, contact staff. (Use the specification fields in this online form for the plain-language description of the specifications)
This is not an eMeasure **Attachment:**

S.2b. **Data Dictionary, Code Table, or Value Sets** *(and risk model codes and coefficients when applicable) must be attached. (Excel or csv file in the suggested format preferred - if not, contact staff)*
No data dictionary **Attachment:**

S.3. **For endorsement maintenance**, please briefly describe any changes to the measure specifications since last endorsement date and explain the reasons.
The current submission for maintenance of endorsement for the elderly surgery outcomes measure removes venous thromboembolic events (VTE), which include both deep venous thrombosis and pulmonary embolism, from the measure. This change was prompted by recent publications demonstrating that VTE is highly subject to surveillance bias. A study of 2,838 hospitals found that increased VTE prophylaxis adherence was associated with worse risk-adjusted VTE event rates.(Bilimoria, Chung et al. 2013) Paradoxically hospitals with higher quality, identified by number of accreditations and quality initiatives, had worse VTE rates. The explanation for this paradoxical relationship is suggested by the association of higher rates of VTE imaging studies among these hospitals with higher rates of VTE detection. (Bilimoria, Chung et al. 2013, Ju, Chung et al. 2014, Chung, Ju et al. 2015)

Details concerning measure performance with and without inclusion of VTE are included in the Data Testing supplement. The inclusion of socioeconomic status (SES) data has also been evaluated in the Data Testing supplement. Measure performance with the inclusion of SES data was evaluated with three additional variables: race, Hispanic ethnicity, and income. These variables did not significantly change or improve the measure performance and, therefore, have not been added to the measure specifications.

Bilimoria, K. Y., J. Chung, M. H. Ju, E. R. Haut, D. J. Bentrem, C. Y. Ko and D. W. Baker (2013). "Evaluation of surveillance bias and the validity of the venous thromboembolism quality measure." Jama 310(14): 1482-1489.
Chung, J. W., M. H. Ju, C. V. Kinnier, M. W. Sohn and K. Y. Bilimoria (2015). "Postoperative venous thromboembolism outcomes measure: analytic exploration of potential misclassification of hospital quality due to surveillance bias." Ann Surg 261(3): 443-444.
Ju, M. H., J. W. Chung, C. V. Kinnier, D. J. Bentrem, D. M. Mahvi, C. Y. Ko and K. Y. Bilimoria (2014). "Association between hospital imaging use and venous thromboembolism events rates based on clinical data." Ann Surg 260(3): 558-564; discussion 564-556.

S.4. **Numerator Statement** *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome)*
IF an OUTCOME MEASURE, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.
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.

S.5. **Time Period for Data** *(What is the time period in which data will be aggregated for the measure, e.g., 12 mo, 3 years, look back to August for flu vaccination? Note if there are different time periods for the numerator and denominator.)*
Targeted events within 30 days of the operation are included.

S.6. **Numerator Details** *(All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, specific data collection items/responses, code/value sets – Note: lists of*

individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)
IF an OUTCOME MEASURE, describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in the calculation algorithm.

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.

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. original index hospital or at an outside hospital).

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.

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 greater than 3 times upper level of the reference range in the setting of suspected myocardial ischemia

c. Physician diagnosis of myocardial infarction.

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.

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 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.

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.

Wound Disruption: Separation of the layers of a surgical wound, which may be partial or complete, with disruption of the 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.

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 from minimally contaminated lower respiratory tract (LRT) specimen (e.g. BAL or protected

specimen brushing)

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 from preoperative value, but with no requirement for dialysis.

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 five:
a. fever (>38 degrees C),
b. urgency,
c. frequency,
d. dysuria,
e. suprapubic 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
AND ANY ONE or MORE of the following seven:
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.

S.7. Denominator Statement *(Brief, narrative description of the target population being measured)*
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)

S.8. Target Population Category *(Check all the populations for which the measure is specified and tested if any):*
Senior Care

S.9. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, specific data collection items/responses , code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should*

be provided in an Excel or csv file in required format at S.2b)
Cases are collected so as to match ACS NSQIP inclusion and exclusion criteria, thereby permitting valid application of ACS NSQIP model-based risk adjustment.

S.10. **Denominator Exclusions** *(Brief narrative description of exclusions from the target population)*
Cases must first have ACS NSQIP eligible CPT codes on the submitted list of ~2900 codes. Major/multisystem trauma and transplant surgeries are excluded. Patients who are ASA 6 (brain-death organ donor) are not eligible surgical cases. Surgeries following within 30 d of an index procedure are an outcome (return to OR) and are not eligible to be new index cases. Thus, a patient known to have had a prior surgical operation within 30 days is excluded from having the subsequent surgery considered an index case.

S.11. **Denominator Exclusion Details** *(All information required to identify and calculate exclusions from the denominator such as definitions, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b)*
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.

S.12. **Stratification Details/Variables** *(All information required to stratify the measure results including the stratification variables, definitions, 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 with at S.2b)*
The measure is risk adjusted and case mix adjusted.

S.13. **Risk Adjustment Type** (Select type. Provide specifications for risk stratification in S.12 and for statistical model in S.14-15)
Statistical risk model
If other:

S.14. **Identify the statistical risk model method and variables** *(Name the statistical method - e.g., logistic regression and list all the risk factor variables. Note - risk model development and testing should be addressed with measure testing under Scientific Acceptability)*
ACS NSQIP performs hospital-level profiling by reporting case-mix adjusted and risk-adjusted postoperative outcomes. The statistical modeling is performed in three steps, which include case-mix adjustment, variable selection, then risk adjustment, all of which are carried out using the SAS software package (v 9.2).

In the first step, clinically similar procedures (defined by CPT codes) are categorized into established groups. Generalized linear mixed modeling (GLMM, also called hierarchical modeling in this measure) is used to calculate linear predictor values for each procedure group (SAS PROC GLIMMIX). These linear predictors (referred to as "CPT Risk") rank each procedure group on a continuous scale based on the log probability for outcome, and are risk adjusted for patient factors. The CPT Risk variable provides case-mix adjustment for the hospital profiling.

For variable selection of risk factors, step-wise logistic regression (SAS PROC LOGISTIC) is performed using NSQIP predictors. The NSQIP predictors demonstrating statistical significance ($P<0.05$) are selected for the preliminary predictor list. A subset of this list is chosen based on clinical relevance, statistical importance, and ease of data extraction to create a small, fixed or "parsimonious" predictor set. This composite mortality or any serious morbidity outcome measure was evaluated based on the following three predictors: ASA class, CPT risk and functional status.

In the final step, both case-mix adjustment and risk adjustment are performed for the hospital profiling using the CPT Risk and the parsimonious predictor set, respectively. A GLMM is created (SAS PROC GLIMMIX) which reflects the hierarchical nature of the data,

with patients clustered within hospitals (random intercept, fixed slope model with logistic regression). The model incorporates the empirical Bayes method, which optimally combines information from the particular hospital with information from the sample of all hospitals to arrive at a best prediction about each hospital's performance. Sometimes called a reliability adjustment, but more properly described as smoothing or pooling, this adjustment tends to shrink predicted hospital performance towards the grand mean hospital value, with the effect of shrinkage greatest when the hospital sample size is small and when the hospital's estimate is extreme compared to other hospitals.

Hospital performance is reported as an odds ratio (the odds for the hospital versus the odds for the statistically constructed average hospital). Hospitals with odds ratios less than 1.0 demonstrate better than average performance; those with odds ratios greater than 1.0 demonstrate worse than average performance. Odds ratios are reported with 95% confidence intervals: the interval does not overlap 1.0, the hospital is designated as a statistically significant high or low outlier, depending on whether the interval is entirely above or below 1.0, respectively.

An outcome was defined as 30-day mortality or any serious morbidity including: cardiac arrest requiring CPR, myocardial infarction, sepsis, septic shock, organ space SSI, deep incisional 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, urinary tract infection, or return to the operating room, according to ACS NSQIP definitions.

Reliability is used to evaluate the hospital profiling; this metric describes how confidently the performance of one hospital can be distinguished from other hospitals. Reliability was assessed using a standard method (described in: Huffman, Cohen et al. 2015), which uses information provided by a random intercept, fixed slope, hierarchical model (implemented by SAS PROC GLIMMIX). Please see Measure Testing attachment.

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

S.15. Detailed risk model specifications *(must be in attached data dictionary/code list Excel or csv file. Also indicate if available at measure-specific URL identified in S.1.)*
Note: Risk model details (including coefficients, equations, codes with descriptors, definitions), should be provided on a separate worksheet in the suggested format in the Excel or csv file with data dictionary/code lists at S.2b.
Provided in response box S.15a

S.15a. Detailed risk model specifications (*if not provided in excel or csv file at S.2b*)
See Measure Testing attachment.

S.16. Type of score:
Ratio
If other:

S.17. Interpretation of Score *(Classifies interpretation of score according to whether better quality is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score)*
Better quality = Lower score

S.18. Calculation Algorithm/Measure Logic *(Describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; aggregating data; risk adjustment; etc.)*
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.

S.19. Calculation Algorithm/Measure Logic Diagram URL or Attachment *(You also may provide a diagram of the Calculation Algorithm/Measure Logic described above at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

S.20. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*
IF a PRO-PM, identify whether (and how) proxy responses are allowed.

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.

S.21. **Survey/Patient-reported data** *(If measure is based on a survey, provide instructions for conducting the survey and guidance on minimum response rate.)*
IF a PRO-PM, specify calculation of response rates to be reported with performance measure results.

S.22. **Missing data** (specify how missing data are handled, e.g., imputation, delete case.)
Required for Composites and PRO-PMs.
ACS NSQIP has placed a very high value on accuracy of data collection while maintaining a sample size large enough for statistical modeling and keeping within regulations for patient confidentiality. The methodology of our program has been highly successful with increasing numbers of participants every year, and measureable improvements in surgical outcomes over time based on the O/E ratios for mortality and various post-surgical complications. Historically, the use of trained data collectors within ACS NSQIP and a comprehensive support system has resulted in high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Data definitions are continually evaluated and inter-rater reliability audits are regularly performed.

S.23. **Data Source** *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*
If other, please describe in S.24.
Electronic Clinical Data, Electronic Clinical Data : Electronic Health Record, Electronic Clinical Data : Imaging/Diagnostic Study, Electronic Clinical Data : Laboratory, Electronic Clinical Data : Pharmacy, Electronic Clinical Data : Registry, Management Data, Paper Medical Records

S.24. **Data Source or Collection Instrument** *(Identify the specific data source/data collection instrument e.g. name of database, clinical registry, collection instrument, etc.)*
IF a PRO-PM, identify the specific PROM(s); and standard methods, modes, and languages of administration.
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.

S.25. **Data Source or Collection Instrument** *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*
URL

S.26. **Level of Analysis** *(Check ONLY the levels of analysis for which the measure is SPECIFIED AND TESTED)*
Facility

S.27. **Care Setting** *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*
Ambulatory Care : Ambulatory Surgery Center (ASC), Hospital/Acute Care Facility
If other:

S.28. **COMPOSITE Performance Measure** - Additional Specifications *(Use this section as needed for aggregation and weighting rules, or calculation of individual performance measures if not individually endorsed.)*

2a. Reliability – See attached Measure Testing Submission Form
2b. Validity – See attached Measure Testing Submission Form
0697_MeasureTesting_Maintenance-May2016.doc

NQF #: 0697 **NQF Project:** Patient Outcomes Measures: Phases I and II

2. RELIABILITY & VALIDITY - SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented. **(evaluation criteria)**
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 field. Supplemental materials may be referenced or attached in item 2.1. See guidance on measure testing.

2a2. Reliability Testing. (*Reliability testing was conducted with appropriate method, scope, and adequate demonstration of reliability.*)

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.

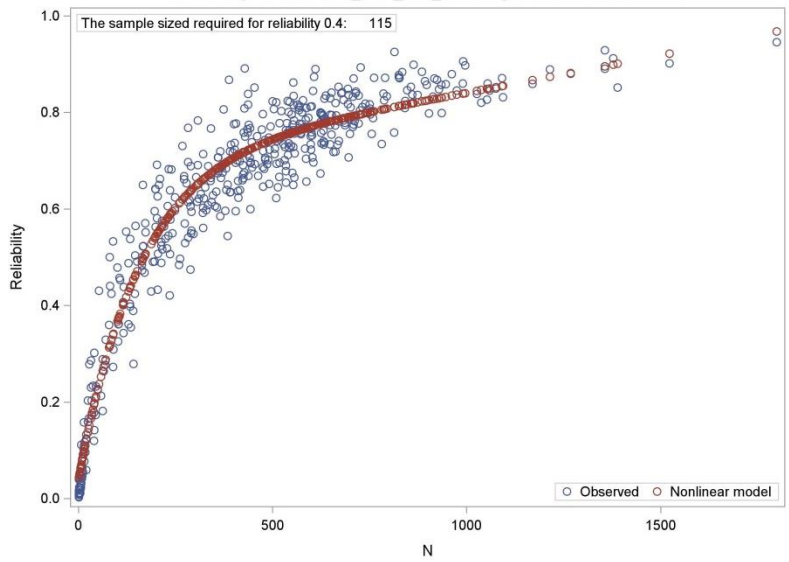


Figure: Reliability vs Case number for Elderly Surgery measure (without VTE and without SES), regression and observed

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

2b. VALIDITY. Validity, Testing, including all Threats to Validity: H☐ M☐ L☐ I ☐

2b1.1 Describe how the measure specifications *(measure focus, target population, and exclusions)* **are consistent with the evidence cited in support of the measure focus (***criterion 1c)* **and identify any differences from the evidence:**

The importance of tracking morbidity and mortality for elderly individuals undergoing surgery is emphasized in the evidence and consistent with the current measure specifications. Please see criterion 1c for additional information.

2b2. Validity Testing. (*Validity testing was conducted with appropriate method, scope, and adequate demonstration of validity.)*

2b2.1 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 sample derived from the ACS NSQIP database for 2011-2014. See 2a1 above.

2b2.2 Analytic Method *(Describe method of validity testing and rationale; if face validity, describe systematic assessment)***:**
See Risk-adjustment Methodology in Specifications.

May 31, 2016 Maintenance of Endorsement Update:

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.

2b2.3 Testing Results *(Statistical results, assessment of adequacy in the context of norms for the test conducted; if face validity, describe results of systematic assessment)*:
See Risk-adjustment Methodology in Specifications. Model validity (a similar c-statistic, discrimination) was demonstrated when the 2008 model was applied to 2007 data.

May 31, 2016 Maintenance of Endorsement Update:

See Risk-adjustment Methodology in Specifications.

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

POTENTIAL THREATS TO VALIDITY. (*All potential threats to validity were appropriately tested with adequate results.*)

2b3. Measure Exclusions. (*Exclusions were supported by the clinical evidence in 1c or appropriately tested with results demonstrating the need to specify them.*)

2b3.1 **Data/Sample for analysis of exclusions** *(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 2011-2014. See 2a1 above.

2b3.2 **Analytic Method** (*Describe type of analysis and rationale for examining exclusions, including exclusion related to patient preference)***:**
See Risk-adjustment Methodology in Specifications.

2b3.3 **Results** *(Provide statistical results for analysis of exclusions, e.g., frequency, variability, sensitivity analyses):*
See Risk-adjustment Methodology in Specifications.

2b4. Risk Adjustment Strategy. (*For outcome measures, adjustment for differences in case mix (severity) across measured entities was appropriately tested with adequate results.*)

2b4.1 **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)***:**

The data sample is derived from the most recent ACS NSQIP Data file . The model for patients >= 65 used 83,832 patient records. Future models can be constructed using the most recent Data file. If this measure is adopted by sufficient numbers of non-NSQIP hospitals re-modeling can be based on data from the broader sample of hospitals.

May 31, 2016 Maintenance of Endorsement Update:
The data sample of 655,187 patients age 65 and older was included from ACS NSQIP Data files (2011-2014) for model generation.

2b4.2 **Analytic Method (***Describe methods and rationale for development and testing of risk model or risk stratification including selection of factors/variables)***:**
Preliminary risk-adjustment models were constructed for these developmental purposes using step-wise logistic regression, and GEE models were also utilized which allow for effect of clustering. Compared to full hierarchical models this methodology poses fewer convergence problems, has step-wise variable-selection methodology, and we have found that it provides nearly identical risk-adjustment as random intercept hierarchical models. However, Odds ratios and parameters reported here are derived from hierarchical model methodology applied to the predictor set established using step-wise logistic regression methods.

May 31, 2016 Maintenance of Endorsement Update:

It was our intention to estimate model parameter values (more accurately) from a large, multi-year sample and then apply historical prediction equations to samples composed of annual accumulations of data. This required a logistic rather than a hierarchical approach (which involves contemporaneous data modeling), with the quality metric being an O/E ratio. Step-wise logistic regression, informed by clinical insights and the need for parsimony, resulted in prediction equations with either 3 factors (CPT category, ASA Class, Preoperative Functional Status (Independent, Partially Dependent, Totally Dependent)) or 6 (with 3 additional variables for exploration of SES: Median Income, Hispanic Ethnicity, Race).

We examined differences in risk-adjusted outcomes when equations were applied to the 2014 data. Specifically, when the SES variables were or were not included, and when VTE was or was not included as an outcome (thus, there were 4 sets of parameter values). SES was examined to determine whether this factor would represent a crucial risk-adjustment component of quality, and VTE was examined as there is evidence that the observation of a VTE event is subject to a substantial surveillance bias such that it is an inappropriate outcome for quality monitoring – "good" hospitals that initiate careful, sometimes universal, surveillance are at a

disadvantage compared to other hospitals that are less likely to look for non-symptomatic VTEs.

Hierarchical modeling provides several theoretical advantages over ordinary logistic modeling including appropriate consideration of the nested structure of data (patients within hospitals) and the automatic incorporation of an empirical-Bayes-type shrinkage adjustment to stabilize estimates (of particular importance when sample sizes are small and event rates low). Our own research has indicated the adjustment of error variance estimates associated with nesting has little practical effect. However, shrinkage adjustments do provide, under certain conditions, for better quality estimates, although shrinkage can potentially mask real quality differences (i.e., the approach can be overly conservative). While not reported on in this submission, we are exploring the incorporation of post-logistic modeling smoothing (shrinkage) to Measure Elderly Surgery O/E ratios. This methodology, as applied generally to ACS NSQP data, has been described elsewhere (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*, in press.)

2b4.3 Testing Results *(Statistical risk model: Provide quantitative assessment of relative contribution of model risk factors; risk model performance metrics including cross-validation discrimination and calibration statistics, calibration curve and risk decile plot, and assessment of adequacy in the context of norms for risk models. Risk stratification: Provide quantitative assessment of relationship of risk factors to the outcome and differences in outcomes among the strata)*:
See Risk-adjustment Methodology in Specifications. 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.

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.

2b4.4 **If outcome or resource use measure is not risk adjusted, provide rationale and analyses to justify lack of adjustment:**
Risk adjusted

2b5. Identification of Meaningful Differences in Performance. (*The performance measure scores were appropriately analyzed and discriminated meaningful differences in quality.*)

2b5.1 **Data/Sample** *(Describe 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 Strategy Data Sample Section. (2b4.1).

2b5.2 **Analytic Method** *(Describe methods and rationale to identify statistically significant and practically/meaningfully differences in performance)***:**
The default methodology for discrimination performance is based on the computed 95% CI (using Ulm's method) for the O/E ratio. If the interval is entirely above1.0, the hospital is identified as having performance significantly worse than expected. If the interval is entirely below 1.0, the hospital is identified as having performance significantly better than expected. If the interval overlaps 1.0 the hospital is performing "as expected." Depending on programmatic objectives, the implementing organization could also opt for outlier status being defined by upper and lower percentile ranks in O/E ratios

2b5.3 **Results** *(Provide measure performance results/scores, e.g., distribution by quartile, mean, median, SD, etc.; identification of statistically significant and meaningfully differences in performance)*:
See Risk-adjustment strategy Testing Results (2b4.3)

2b6. Comparability of Multiple Data Sources/Methods. (*If specified for more than one data source, the various approaches result in comparable scores.*)

2b6.1 **Data/Sample** *(Describe the data or sample including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included)***:**
The only sources of data are those indicated above. This measure requires clinical data (electronic or paper records), with administrative data added only as necessary.

May 31, 2016 Maintenance of Endorsement Update:

The current maintenance of endorsement submission provides measure performance with the addition of the following SES data: race, Hispanic ethnicity, and income, as estimated by proxy using median income for patient zip code mapped via the University of Michigan Population Studies Center zip code characteristics (available at http://www.psc.isr.umich.edu/dis/census/Features/tract2zip/)

The advantage of clinical data versus administrative or claims data in identifying risk-adjusted outcomes is exemplified in the study by Steinberg et al (2008). The study compared comorbidities collected and postsurgical complications from the ACS NSQIP database and the University HealthSystem Consortium (UHC). Comorbidities per patient were identified twice as often in the UHC system, while there was a discordance of 26% in identifying complications (UHC complication rate, 2% vs. ACS NSQIP complication rate, 28%). Recent studies have compared ACS NSQIP data and Medicare claims data, indicating lack of agreement and poor correlation between the two data sources as it relates to complication identification and risk-adjustment.(Lawson, Zingmond et al. 2015, Lawson, Louie et al. 2016) Using administrative or claims data may result in significant differences in risk-adjusted outcomes than using clinical data.

Lawson, E. H., R. Louie, D. S. Zingmond, G. D. Sacks, R. H. Brook, B. L. Hall and C. Y. Ko (2016). "Using Both Clinical Registry and Administrative Claims Data to Measure Risk-adjusted Surgical Outcomes." Ann Surg 263(1): 50-57.
Lawson, E. H., D. S. Zingmond, B. L. Hall, R. Louie, R. H. Brook and C. Y. Ko (2015). "Comparison between clinical registry and medicare claims data on the classification of hospital quality of surgical care." Ann Surg 261(2): 290-296.
Steinberg, S.M., et al., Comparison of risk adjustment methodologies in surgical quality improvement. Surgery, 2008. 144(4): p. 662-7; discussion 662-7.

2b6.2 Analytic Method *(Describe methods and rationale for testing comparability of scores produced by the different data sources specified in the measure)*:
See above

2b6.3 Testing Results *(Provide statistical results, e.g., correlation statistics, comparison of rankings; assessment of adequacy in the context of norms for the test conducted)*:
See above

2c. Disparities in Care: H☐ M☐ L☐ I ☐ NA☐ (*If applicable, the measure specifications allow identification of disparities.*)

2c.1 If measure is stratified for disparities, provide stratified results *(Scores by stratified categories/cohorts)*: measure is not stratified; measure is case mix adjusted, without inclusion of race or ethnicity. Post hoc stratification by race or ethnicity can be performed for the purpose of identifying disparities.

May 31, 2016 Maintenance of Endorsement Update:
As mentioned above, the current submission includes measure testing with SES data, including race, ethnicity and income (estimated using zip code, as described above). Please see Testing Results (in particular 2a2.3, 2b2.3 and 2b4.3) for additional details regarding model performance when SES data is included.

2c.2 If disparities have been reported/identified (e.g., in 1b), but measure is not specified to detect disparities, please explain:
N/A

2.1-2.3 Supplemental Testing Methodology Information:

Steering Committee: Overall, was the criterion, *Scientific Acceptability of Measure Properties*, met?
(*Reliability and Validity must be rated moderate or high*) Yes☐ No☐
Provide rationale based on specific subcriteria:

If the Committee votes No, STOP

3. Feasibility

Extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

3a. Byproduct of Care Processes

For clinical measures, the required data elements are routinely generated and used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order).

3a.1. Data Elements Generated as Byproduct of Care Processes.
Data generated as byproduct of care processes during care delivery (Data are generated and used by healthcare personnel during the provision of care, e.g., blood pressure, lab value, medical condition), Coding/abstraction performed by someone other than person obtaining original information (E.g., DRG, ICD-9 codes on claims, chart abstraction for quality measure or registry)
If other:

3b. Electronic Sources

The required data elements are available in electronic health records or other electronic sources. If the required data are not in electronic health records or existing electronic sources, a credible, near-term path to electronic collection is specified.

3b.1. **To what extent are the specified data elements available electronically in defined fields?** *(i.e., data elements that are needed to compute the performance measure score are in defined, computer-readable fields)*
No data elements are in defined fields in electronic sources

3b.2. **If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using other than electronic sources.**
A completely electronic medical record (EMR) will be needed to capture the risk factors that enter into the model. In addition, a software module (currently available to ACS NSQIP subscribers) will be required to transfer information from the EMR to a measure submission database. ACS NSQIP is in the process of developing automated data extraction from EMR vendors, however, electronic data entry is currently not available for clinical information required in the elderly surgery outcome measure.

3b.3. **If this is an eMeasure, provide a summary of the feasibility assessment in an attached file or make available at a measure-specific URL.**
Attachment:

3c. Data Collection Strategy

Demonstration that the data collection strategy (e.g., source, timing, frequency, sampling, patient confidentiality, costs associated with fees/licensing of proprietary measures) can be implemented (e.g., already in operational use, or testing demonstrates that it is ready to put into operational use). For eMeasures, a feasibility assessment addresses the data elements and measure logic and demonstrates the eMeasure can be implemented or feasibility concerns can be adequately addressed.

3c.1. **Describe what you have learned/modified as a result of testing and/or operational use of the measure regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.**
IF a PRO-PM, consider implications for both individuals providing PROM data (patients, service recipients, respondents) and those whose performance is being measured.
ACS NSQIP has been open to subscription by private sector hospitals since 2004. Ten years prior to this time the program was implemented in the U.S. Department of Veterans Affairs. Thus we have long term experience with the data collection and operational use of the O/E ratio for quality improvement and benchmarking on which this measure is based. Historically, the use of trained data collectors within ACS NSQIP and a comprehensive support system has resulted in high reliability of data and very few problems with missing data. Participants in the program are required to assign a dedicated person for data collection to ensure reliable assessment of clinical data.

Data definitions are continually evaluated and inter-rater reliability audits are regularly performed.

ACS NSQIP has placed a very high value on accuracy of data collection while maintaining a sample size large enough for statistical modeling and keeping within regulations for patient confidentiality. The methodology of our program has been highly successful with increasing numbers of participants every year, and measureable improvements in surgical outcomes over time based on the O/E ratios for mortality and various post-surgical complications. Due to the much smaller number of variables needed for participation in this measure than in the full program, we expect that hospitals that are not ACS NSQIP participants will also be able to achieve highly reliable results.

3c.2. **Describe any fees, licensing, or other requirements to use any aspect of the measure as specified** *(e.g., value/code set, risk model, programming code, algorithm)*.

4. Usability and Use

Extent to which potential audiences (e.g., consumers, purchasers, providers, policy makers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

4a. Accountability and Transparency

Performance results are used in at least one accountability application within three years after initial endorsement and are

publicly reported within six years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then a credible plan for implementation within the specified timeframes is provided.

4.1. Current and Planned Use

NQF-endorsed measures are expected to be used in at least one accountability application within 3 years and publicly reported within 6 years of initial endorsement in addition to performance improvement.

Planned	Current Use (for current use provide URL)
	Public Reporting Hospital Compare https://www.medicare.gov/hospitalcompare/acs-surgical-measures.html Quality Improvement with Benchmarking (external benchmarking to multiple organizations) ACS NSQIP https://www.facs.org/quality-programs/acs-nsqip Quality Improvement (Internal to the specific organization) ACS NSQIP https://www.facs.org/quality-programs/acs-nsqip

4a.1. For each CURRENT use, checked above, provide:

- Name of program and sponsor
- Purpose
- Geographic area and number and percentage of accountable entities and patients included

Public Reporting: The Elderly surgery outcomes measure is reported in Hospital Compare, through the Centers for Medicare & Medicaid Services website. The purpose of the website is to provide the public with freely accessible quality information with which to compare hospitals. According to the data reporting period for April 2016, there are 131 hospitals currently reporting their risk-adjusted surgery outcomes data for NQF-endorsed measures from ACS. Please see https://www.medicare.gov/hospitalcompare/acs-surgical-measures.html for detailed list. (Accessed 5/16/2016)

Quality Improvement, both internal and external with benchmarking: The program ACS NSQIP is a quality improvement registry that provides hospitals with clinical data with which to track and improve outcomes, as well as benchmarked, risk-adjusted outcomes reports. The Elderly surgery outcomes measure is currently provided to NSQIP hospitals. There are over 600 hospitals across the U.S. currently participating in ACS NSQIP and receiving risk-adjusted benchmarking reports. Hospitals utilize their internal data for the purpose of quality improvement initiatives specific to the organization.

4a.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)

NA

4a.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

NA

4b. Improvement

Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

4b.1. Progress on Improvement. (Not required for initial endorsement unless available.)

Performance results on this measure (current and over time) should be provided in 1b.2 and 1b.4. Discuss:

- **Progress (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**

A recent analysis indicates that over 8 years in the program, 62% and 71% of hospitals improve their performance in mortality and risk-adjusted complications. (Cohen et al. 2016) Annual reductions are approximately 0.8% in mortality and 3.1% in morbidity; though small, these reductions provide cumulative benefit as hospitals continue participation in the ACS NSQIP program. For 2014, there were 460 hospitals contributing 206,064 surgeries on adults age 65 and older. The O/E ratios for elderly surgery mortality and serious morbidity range from 0.55 to 1.6 for participating hospitals. The interquartile range for the O/E ratio is 0.22, and the 10th percentile and 90th percentile O/E ratios were 0.82 and 1.22, respectively. These numbers indicate that although there have been improvements over time, a performance gap remains between those performing better and worse than expected after risk and case mix adjustment.

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

4b.2. **If no improvement was demonstrated, what are the reasons? If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.**

4c. Unintended Consequences

The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).

4c.1. **Were any unintended negative consequences to individuals or populations identified during testing; OR has evidence of unintended negative consequences to individuals or populations been reported since implementation? If so, identify the negative unintended consequences and describe how benefits outweigh them or actions taken to mitigate them.**

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.

5. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

Are there related measures (conceptually, either same measure focus or target population) or competing measures (conceptually both the same measure focus and same target population)? If yes, list the NQF # and title of all related and/or competing measures.

Yes

5.1a. **List of related or competing measures (selected from NQF-endorsed measures)**

0534 : Hospital specific risk-adjusted measure of mortality or one or more major complications within 30 days of a lower extremity bypass (LEB).

0706 : Risk Adjusted Colon Surgery Outcome Measure

5.1b. **If related or competing measures are not NQF endorsed please indicate measure title and steward.**

5a. Harmonization

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. **If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed**

measure(s):
Are the measure specifications completely harmonized?
Yes
5a.2. **If the measure specifications are not completely harmonized, identify the differences, rationale, and impact on interpretability and data collection burden.**

5b. Competing Measures

The measure is superior to competing measures (e.g., is a more valid or efficient way to measure);
OR
Multiple measures are justified.

5b.1. **If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):**
Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)
NA - different target populations

Appendix

A.1 Supplemental materials may be provided in an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be organized in one file with a table of contents or bookmarks. If material pertains to a specific submission form number, that should be indicated. Requested information should be provided in the submission form and required attachments. There is no guarantee that supplemental materials will be reviewed.
Attachment **Attachment:** Appendix_CPT_Code_Inclusion_List.pdf

Contact Information

Co.1 **Measure Steward (Intellectual Property Owner):** American College of Surgeons
Co.2 **Point of Contact:** Sameera, Ali, sali@facs.org, 312-202-5431-
Co.3 **Measure Developer if different from Measure Steward:** American College of Surgeons
Co.4 **Point of Contact:** Sameera, Ali, sali@facs.org, 312-202-5431-

Additional Information

Ad.1 **Workgroup/Expert Panel involved in measure development**
Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.
American College of Surgeons, Division of Research and Optimal Patient Care, Section of Continuous Quality Improvement
Clifford Ko
Sameera Ali
Bruce Hall
Mark Cohen
Yaoming Liu
Julia Berian

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.

Measure Developer/Steward Updates and Ongoing Maintenance
Ad.2 **Year the measure was first released:** 2011
Ad.3 **Month and Year of most recent revision:** 05, 2016
Ad.4 **What is your frequency for review/update of this measure?** Annual
Ad.5 **When is the next scheduled review/update for this measure?** 05, 2017

Ad.6 **Copyright statement:**
Ad.7 **Disclaimers:**

Ad.8 **Additional Information/Comments:**