NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Surgery Standing Committee – Fall 2020 Measure Evaluation Web Meetings

The National Quality Forum (NQF) convened the Surgery Standing Committee (link to slides) for two web meetings on February 12 and 16, 2021, to evaluate eight maintenance measures.

Welcome, Introductions, and Review of Meeting Objectives

Amy Moyer, NQF director, welcomed the Standing Committee and participants to the web meeting. NQF staff reviewed the agenda. Standing Committee members each introduced themselves and disclosed any conflicts of interest. No Surgery Standing Committee members were recused due to conflicts of interest for any of the eight measures under review for the fall 2020 cycle. One Standing Committee member was recused from voting on the Scientific Acceptability criteria for measures that had been reviewed by the Scientific Methods Panel (SMP). This Standing Committee member was recused because they had already voted on these measures as a member of the SMP.

During the first measure evaluation meeting on February 12, 2021, some Standing Committee members were unable to attend the entire meeting due to early departures and late arrivals. The vote totals reflect members present and eligible to vote. Quorum (14 out of 20 Standing Committee members) was met and maintained for the entirety of this meeting.

During the second measure evaluation meeting on February 16, 2021, voting quorum was not achieved. Therefore, the Standing Committee discussed all relevant criteria and voted after the meeting using an online voting tool.

Topic Area Introduction and Overview of Evaluation Process

NQF staff provided an overview of the topic area and the current NQF portfolio of endorsed measures. There are currently 58 endorsed measures in the Surgery portfolio. Additionally, NQF reviewed the Consensus Development Process (CDP) and the measure evaluation criteria. A measure is recommended for endorsement by the Standing Committee when the vote margin on all must-pass criteria (Importance, Scientific Acceptability, Use), and overall, is greater than 60 percent of voting members in favor of endorsement. A measure is not recommended for endorsement when the vote margin on any must-pass criterion or overall is less than 40 percent of voting members in favor of endorsement. The Standing Committee has not reached consensus if the vote margin on any must-pass criterion or overall is between 40 and 60 percent, inclusive, in favor of endorsement. When the Standing Committee has not reached consensus, all measures for which consensus was not reached will be released for NQF member and public comment. The Standing Committee will consider the comments and re-vote on those measures during a webinar convened after the commenting period closes.

Measure Evaluation

During the meeting, the Surgery Standing Committee evaluated eight maintenance measures for endorsement consideration. NQF solicits comments for four weeks prior to the measure evaluation meeting. For this evaluation cycle, the commenting period opened on December 23, 2020. Ten

comments were submitted by the pre-meeting deadline (January 26, 2021) and shared with the Standing Committee prior to the measure evaluation meeting. Those comments are included at the end of this summary. A summary of the Standing Committee's deliberations will be compiled and provided in the draft technical report. NQF will post the draft technical report on April 1, 2021, for public comment on the NQF website. The draft technical report will be posted for 30 calendar days.

Rating Scale: H – High; M – Medium; L – Low; I – Insufficient; NA – Not Applicable

NQF #0117 Beta Blockade at Discharge (The Society of Thoracic Surgeons (STS))

Description: Percent of patients aged 18 years and older undergoing isolated CABG who were discharged on beta blockers; **Measure Type**: Process; **Level of Analysis**: Facility, Clinician: Group/Practice; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS
Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: H-0; M-18; L-0; I-0 (denominator = 18)
- Performance Gap: H-2; M-4; L-12; I-0 (denominator = 18)
- Reliability: H-1; M-15; L-1; I-0 (denominator = 17)
- Validity: H-2; M-11; L-2; I-2 (denominator = 17)
- Feasibility: H-7; M-10; L-1; I-0 (denominator = 18)
- Use: Pass-18; No Pass-0 (denominator = 18)
- Usability: H-7; M-10; L-1; I-0 (denominator = 18)

Standing Committee Recommendation for Inactive Endorsement with Reserve Status: Yes-17; No-0 (Denominator = 17)

The Standing Committee recommended the measure for inactive endorsement with reserve status.

This measure was discussed during the first measure evaluation web meeting. The Standing Committee noted that this process measure is part of the "use of all evidence-based perioperative medications" domain in NQF #0696 *STS CABG Composite Score*. Originally endorsed in 2007, and most recently endorsed in 2017, this measure focuses on patients discharged on a beta blocker following coronary artery bypass graft (CABG) surgery. The Standing Committee noted that the evidence was largely unchanged from the previous maintenance cycle. A Standing Committee member mentioned that a large new study was recently published this year (2021) that strengthens the existing evidence for postoperative use of beta blockers.

The Standing Committee and developers engaged in a robust conversation about what constitutes a meaningful performance gap and the implications of placing a measure on reserve status. The Standing Committee noted that the performance appears fairly topped out, with median rates of 100 percent and

little variation by insurance type, gender, or race. Standing Committee members shared that with performance rates this high, a great deal of resources are required to achieve a small gain and those resources may be better spent on more impactful areas. A Standing Committee member raised a concern that when the overall performance is this high, a participant needs to perform perfectly to score well. Another Standing Committee member raised a concern regarding whether performance would remain high if the measure were to be placed on reserve status. The developer echoed this concern, adding that they view cardiothoracic surgery as the ultimate high-reliability surgery and that all participants should achieve 100 percent on this measure. They also clarified that they do not penalize small volume programs unless there is a statistically significant gap in performance. The developer also stated that they will continue to collect and use this measure; therefore, the benefit to reserve status may be limited. The Standing Committee voted and reached consensus that the measure did not have a sufficient performance gap to warrant maintaining active endorsement. Ms. Moyer described the process, criteria, and rationale for reserve status. When improvement in performance on an endorsed measure has closed the performance gap and the measure continues to meet all other endorsement criteria, the Standing Committee can recommend that the measure remain endorsed with reserve status. Reserve status results in measures maintaining endorsement, thereby remaining in the measure portfolio while indicating that the measure may not have a sufficient gap to make it a priority for adoption. The Standing Committee agreed that reserve status should be considered for this measure and continued discussing and voting on the remaining criteria.

The Standing Committee revisited the question of how reliable the measure is for participants with a low sample size. The developer clarified that all STS process measures are binary results (i.e., meets/does not meet) with a confidence interval. In general, the smaller the sample size, the larger the confidence interval, which results in most small groups receiving two stars. A Standing Committee member stated that they appreciated the testing for demonstrating different reliabilities at different case counts, noting that there is a range of reliability for each count. This same Standing Committee member noted that reliability of distribution is helpful and that the reliability of "binning" providers into stars would also be helpful. The Standing Committee was satisfied with the measure's reliability. They had no issues or questions regarding validity.

The Standing Committee held brief discussions related to feasibility and use and usability. They discussed a high rating versus a moderate rating for feasibility, noting that the measure is automatically calculated for providers using the STS Adult Cardiac Registry. Standing Committee members also noted that data submission to the registry requires staff to abstract the data for entry into the registry, and this requirement led to their consideration of feasibility as moderate instead of high. The Standing Committee questioned whether public reporting as part of a composite meets the intent of the use criterion. Ms. Moyer shared that the Standing Committee had previously discussed this matter at length and at that time, they had concluded that this did meet the use criterion. The Standing Committee agreed with this previous conclusion. The Standing Committee raised no questions regarding the usability of the measure and voted unanimously to recommend inactive endorsement with reserve status. Discussion of related measures was deferred to the post-comment web meeting.

NQF #0127 Preoperative Beta Blockade (STS)

Description: Percent of patients aged 18 years and older undergoing isolated CABG who received beta blockers within 24 hours preceding surgery.; **Measure Type**: Process; **Level of Analysis**: Facility, Clinician: Group/Practice; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS

Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: H-0; M-17; L-0; I-0 (denominator = 17)
- Performance Gap: H-1; M-13; L-3; I-0 (denominator = 17)
- Reliability: H-1; M-17; L-0; I-0 (denominator = 18)
- Validity: H-0; M-14; L-3; I-1 (denominator = 18)
- Feasibility: H-6; M-10; L-2; I-0 (denominator = 18)
- Use: Pass-18; No Pass-0 (denominator = 18)
- Usability: H-2; M-15; L-1; I-0 (denominator = 18)

Standing Committee Recommendation for Endorsement: Yes-18; No-0 (Denominator = 18)

The Standing Committee recommended the measure for continued endorsement.

This measure was discussed during the first measure evaluation web meeting. The Standing Committee noted that this process measure is part of the "use of all evidence-based perioperative medications" domain in NQF #0696 *STS CABG Composite Score*. Originally endorsed in 2007, and most recently endorsed in 2017, this measure focuses on patients receiving a beta blocker prior to CABG surgery. The Standing Committee noted that the evidence was unchanged from the previous maintenance cycle. They had no issues regarding the evidence tying this process to patient outcomes.

The Standing Committee noted that while performance on this measure is very high, it is lower than the performance on NQF #0117, with a median rate of 98 percent versus 100 percent for NQF #0117. The Standing Committee discussed whether they would be consistently applying the criteria if they were to vote to pass this measure on performance gap. Standing Committee members pointed out that in addition to NQF #0127 having more overall opportunity for improvement than NQF #0117 at the median, the lower deciles of performance on NQF #0127 also demonstrated greater variability in performance than the lower performance deciles for NQF #0117. The Standing Committee determined that this measure still has enough room for improvement to meet the performance gap criterion.

The Standing Committee noted that the reliability and validity testing methodologies and results were very similar to those used for NQF #0117 and that the same discussion points apply to this measure (NQF #0127). The Standing Committee had no concerns related to feasibility or use and usability and determined that the measure met all of these criteria. Discussion of related measures was deferred to the post-comment web meeting.

NQF #0134 Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG) (STS)

Description: Percentage of patients aged 18 years and older undergoing isolated coronary artery bypass graft (CABG) who received an internal mammary artery (IMA) graft; **Measure Type**: Process; **Level of Analysis**: Facility, Clinician: Group/Practice; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS
Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: H-0; M-18; L-0; I-0 (denominator = 18)
- Performance Gap: H-3; M-7; L-8; I-0 (denominator = 18, consensus not reached)
- Reliability: H-6; M-12; L-0; I-0 (denominator = 18)
- Validity: H-2; M-15; L-1; I-0 (denominator = 18)
- Feasibility: H-6; M-11; L-1; I-0 (denominator = 17)
- Use: Pass-18; No Pass-0 (denominator = 18)
- Usability: H-4; M-13; L-1; I-0 (denominator = 18)

Standing Committee Recommendation for Endorsement: Yes-X; No-X (Denominator = X)

The Standing Committee did not vote on the recommendation for endorsement at the meeting because they did not reach consensus on performance gap—a must-pass criterion. The Standing Committee will re-vote on the measure at the post-comment web meeting on June 1, 2021.

This measure was discussed during the first measure evaluation web meeting. The Standing Committee noted that this process measure is a component measure of the composite NQF #0696 *STS CABG Composite Score*. Originally endorsed in 2007, and most recently endorsed in 2017, this measure focuses on use of the internal mammary artery (IMA) during CABG surgery. The Standing Committee noted that the evidence was largely unchanged from the previous maintenance cycle and passed the measure on evidence.

The Standing Committee noted that the performance gap for this measure was very similar to the one for NQF #0117. The developer expressed strong concerns with considering reserve status for this measure, as it is more closely tied to patient mortality and outcomes than NQF #0117. The developer further shared that it is easier and faster for surgeons to perform a CABG using veins for grafts; therefore, this measure is important to encourage use of the IMA. In response to the assertion that performance on the measure is topped out, the developer noted that a 1 percent decrease in performance would represent 1,500 patients with a poorer outcome. A Standing Committee member questioned whether this measure is the only incentive keeping surgeons "honest" about using the proper grafting technique, especially given the existing mortality and complication measures. The developer noted that the existing measures cover a 30-day period, and the impact of the graft choice would not be evident in that time frame. They stated that while most surgeons will continue to do the right thing, some may not. Other Standing Committee members noted that while they agree the measure is important and that there may be a perverse incentive to not use the IMA for grafting, the criterion under discussion is whether there is a sufficient performance gap to warrant continued active endorsement. The Standing Committee and developers raised questions regarding the impact and intent of reserve status: What does it mean? How might it be perceived? Would measures be difficult to find and use? Ms. Moyer clarified that reserve status measures are still endorsed. The reserve status

indicates that performance on the measure is very good with limited room for improvement. Currently, in NQF's measure search tool, all endorsed measures (both active and inactive reserve status) are listed in search results. A reserve status measure appears no different from an actively endorsed measure until a user selects the measure to learn more about it. The Standing Committee was unable to reach consensus regarding performance gap. They will re-vote on this criterion at the post-comment web meeting on June 1, 2021.

The Standing Committee had no issues with reliability beyond those already discussed for NQF #0117. The Standing Committee was satisfied that the measure was reliable. The Standing Committee noted concerns with using known-groups analysis with the measure score and with using test-retest as a methodology for establishing validity. Despite these concerns, the Standing Committee determined that the measure was valid.

The Standing Committee held brief discussions related to feasibility and use and usability, noting that NQF #0117, NQF #0127, and NQF #0134 are all similar with regard to these criteria. Discussion of related measures was deferred to the post-comment web meeting.

NQF #3030 STS Individual Surgeon Composite Measure for Adult Cardiac Surgery (STS)

Description: 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

Measure Type: Composite; **Level of Analysis**: Clinician: Individual; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS
Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: Pass-18; No Pass-0 (denominator = 18)
- Performance Gap: H-2; M-16; L-0; I-0 (denominator = 18)
- Composite – Quality Construct and Rationale: H-5; M-12; L-0; I-0 (denominator = 17)
- Reliability: H-8; M-9; L-0; I-0 (denominator = 17)
- Validity: H-2; M-15; L-0; I-0 (denominator = 17)
- Composite Quality Construct: H-5; M-12; L-0; I-0 (denominator = 17)
- Feasibility: H-4; M-12; L-1; I-0 (denominator = 17)
- Use: Pass-17; No Pass-0 (denominator = 17)
- Usability: H-2; M-14; L-0; I-1 (denominator = 17)

Standing Committee Recommendation for Endorsement: Yes-17; No-0 (Denominator = 17)

The Standing Committee recommended the measure for continued endorsement.

This complex measure was not reviewed by the SMP prior to the measure evaluation meeting because the testing information submitted was unchanged from the previous submission.

This measure was discussed during the first measure evaluation web meeting. The developer introduced the measure, noting that it is an individual surgeon composite measure that encompasses multiple procedures and endpoints. The Standing Committee noted that the evidence was unchanged from the previous maintenance cycle. They had no issues regarding the evidence tying the components of this composite measure to patient outcomes. The Standing Committee also had no issues with the performance gap or the composite construct and rationale.

The Standing Committee noted that the reliability testing methodology (i.e., the Bayesian approach to generate possible values followed by a Monte Carlo simulation to estimate the true values) for this measure was very sophisticated and expressed appreciation for the innovative technique. The Standing Committee expressed concerns with the circular reasoning in the validity testing, which compared performance on the composite component measures to the overall composite score. The developer shared that there are no external comparisons available for this measure. A Standing Committee member asked for the rationale for including race in the clinical risk model. The developer shared that the model fit suffers if race is not included and while the exact mechanism is unclear, they suspect a genetic component is at work that contributes to poorer outcomes for non-White patients. They also shared that they are working on adding geocoding to patient records in the registry to allow for more

exploration of the impact of social risk factors. The Standing Committee was satisfied that the measure met all of the Scientific Acceptability criteria (i.e., reliability, validity, and composite construct).

The Standing Committee expressed no concerns regarding the feasibility or usability of the measure. A Standing Committee member asked for clarification on the use criterion, which requires a maintenance measure to be in an accountability program within three years of its initial endorsement. Ms. Moyer explained that given the STS' strong track record of publicly reporting its measures, staff determined that the plan for publicly reporting the measure this year was highly credible and the measure would be placed in an accountability program soon, likely before the completion of this endorsement cycle. The Standing Committee accepted this rationale and voted to pass the measure on use. Discussion of related measures was deferred to the post-comment web meeting.

NQF #3031 STS Mitral Valve Repair/Replacement (MVRR) Composite Score (STS)

Description: 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

Measure Type: Composite; **Level of Analysis**: Facility, Clinician: Group/Practice; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS
Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: Pass-16; No Pass-0 (denominator = 16)
- Performance Gap: H-1; M-16; L-0; I-0 (denominator = 17)
- Composite – Quality Construct and Rationale: H-3; M-14; L-0; I-0 (denominator = 17)
- Reliability: H-1; M-16; L-0; I-0 (denominator = 17)
- Validity: H-1; M-16; L-0; I-0 (denominator = 17)
- Composite Quality Construct: H-3; M-14; L-0; I-0 (denominator = 17)
- Feasibility: H-6; M-10; L-1; I-0 (denominator = 17)
- Use: Pass-17; No Pass-0 (denominator = 17)
- Usability: H-1; M-16; L-0; I-0 (denominator = 17)

Standing Committee Recommendation for Endorsement: Yes-17; No-0 (Denominator = 17)

The Standing Committee recommended the measure for continued endorsement.

This complex measure was not reviewed by the SMP prior to the measure evaluation meeting because the testing information submitted was unchanged from the previous submission.

This measure was discussed during the first measure evaluation web meeting. The Standing Committee noted that this measure submission is very similar to the submission for NQF #3030. The Standing Committee agreed that the discussion for NQF #3030 applied to NQF #3031 as well and did not need to be repeated. The Standing Committee also noted that the evidence was unchanged from the previous maintenance cycle. They had no issues with the evidence tying the components of this composite measure to patient outcomes. The Standing Committee also had no issues with the performance gap or the composite construct and rationale. The Standing Committee was satisfied that the measure met all of the Scientific Acceptability criteria (i.e., reliability, validity, and composite construct). The Standing Committee expressed no concerns regarding the feasibility or use and usability of the measure. They noted that this measure is publicly reported, clearly meeting the use criterion. Discussion of related measures was deferred to the post-comment web meeting.

NQF #3032 STS Mitral Valve Repair/Replacement (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score (STS)

Description: 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). 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

Measure Type: Composite; **Level of Analysis**: Facility, Clinician : Group/Practice; **Setting of Care**: Inpatient/Hospital; **Data Source**: Registry Data

Measure Steward/Developer Representatives at the Meeting

David M. Shahian, MD – STS
Jeffrey P. Jacobs, MD – STS
Moritz C. Wyler von Ballmoos, MD, PhD, MPH – STS
Sean M. O'Brien, PhD – STS
Mark S. Antman, DDS, MBA – STS
Banu Yagci, MBA – STS

Standing Committee Votes

- Evidence: Pass-17; No Pass-0 (denominator = 17)
- Performance Gap: H-1; M-16; L-0; I-0 (denominator = 17)
- Composite – Quality Construct and Rationale: H-3; M-14; L-0; I-0 (denominator = 17)
- Reliability: H-0; M-16; L-0; I-0 (denominator = 16)
- Validity: H-0; M-16; L-0; I-0 (denominator = 16)
- Composite Quality Construct: H-1; M-16; L-0; I-0 (denominator = 17)
- Feasibility: H-6; M-10; L-1; I-0 (denominator = 17)
- Use: Pass-17; No Pass-0 (denominator = 17)
- Usability: H-1; M-16; L-0; I-0 (denominator = 17)

Standing Committee Recommendation for Endorsement: Yes-17; No-0 (Denominator = 17)

The Standing Committee recommended the measure for continued endorsement.

This complex measure was not reviewed by the SMP prior to the measure evaluation meeting because the testing information submitted was unchanged from the previous submission.

This measure was discussed during the first measure evaluation web meeting. The Standing Committee noted that this measure is identical to NQF #3031, except for the addition of the CABG procedure. The Standing Committee agreed that no additional discussion was warranted and passed the measure on all criteria. Discussion of related measures was deferred to the post-comment web meeting.

NQF #1550 Hospital-Level Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Yale Center for Outcomes Research & Evaluation (CORE)/Centers for Medicare & Medicaid Services (CMS))

Description: 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 age 65 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).; **Measure Type**: Outcome; **Level of Analysis**: Facility; **Setting of Care**: Inpatient/Hospital; **Data Source**: Claims, Enrollment Data

Measure Steward/Developer Representatives at the Meeting

Duwa Amin, MPH – Yale CORE
Darinka Djordjevic, PhD – Yale CORE
Kashika Sahay, PhD, MPH – Yale CORE
Anna Sigler, MPH – Yale CORE
Huihui Yu, PhD – Yale CORE
Sapha Hassan, MPH – Yale CORE
Kristina Gaffney, BS – Yale CORE
Lisa Suter, MD – Yale CORE
Elizabeth Triche, PhD – Yale CORE
Doris Peter, PhD – Yale CORE
James Poyer, MS, MBA – CMS

Standing Committee Votes

- Evidence: Pass-18; No Pass-0 (denominator = 18)
- Performance Gap: H-0; M-18; L-0; I-0 (denominator = 18)
- This measure is deemed as complex and Scientific Acceptability was evaluated by the NQF SMP. The SMP rated the measure as moderate for both reliability and validity.
    - Because voting was conducted after the meeting using an online voting tool, the Standing Committee voted on the criteria rather than on whether to accept the SMP's ratings.
    - The Standing Committee's rating for Reliability: H-0; M-15; L-2; I-0 (denominator = 17, due to SMP member recusal)
    - The Standing Committee's rating for Validity: H-0; M-14; L-3; I-0 (denominator = 17, due to SMP member recusal)

- Feasibility: H-4; M-13; L-1; I-0 (denominator = 18)
- Use: Pass-18; No Pass-0 (denominator = 18)
- Usability: H-1; M-17; L-0; I-0 (denominator = 18)

Standing Committee Recommendation for Endorsement: Yes-17; No-1 (Denominator = 18)

The Standing Committee recommended the measure for continued endorsement.

This measure was discussed during the second measure evaluation web meeting. Since quorum was not met during the meeting, the Standing Committee discussed all criteria and then voted after the meeting using an online voting tool. This measure estimates a hospital-level, risk-standardized complication rate (RSCR) associated with elective primary THA and TKA in Medicare fee-for-service (FFS) beneficiaries who are age 65 and older. Originally endorsed in 2012, and recently endorsed in 2017, this outcome measure is defined as any one of the specified complications occurring during complication-specific time frames after admission.

The Standing Committee noted the evidence was directionally the same yet stronger than the evidence from the previous maintenance submission. The Standing Committee observed that there was an appropriate measure performance gap and did not express any concerns.

The Standing Committee noted that while the reliability testing methods were robust, there are concerns from public commenters regarding the reliability at the lower end of the case counts. A Standing Committee member who also serves on the SMP noted that reliability standards are currently in flux, but generally higher is better. They stated it would be helpful to see the reliability of classification to obtain a better understanding of the risk of misclassification at different case counts. The developer responded by identifying the two types of reliability testing performed (i.e., signal-to-noise and split sample). They noted that misclassification was rare, with most providers classified as no different than average. The developer attributes this to a narrowing of variation in performance as performance improves, use of a 95 percent confidence interval, and the impact of statistical modeling.

The Standing Committee engaged in a robust discussion on validity. They noted that the measure currently only includes inpatient procedures. As THA/TKA procedures shift to outpatient settings, the change in patient mix for inpatient procedures could be a threat to the validity of the measure. A Standing Committee member noted the inclusion group, Medicare FFS, and requested clarification on the included and excluded populations. The developer clarified that Medicare Advantage patients are not included. The developer also noted that one third of Medicare patients are enrolled in Medicare Advantage plans and that they would seek to incorporate those patients in future versions of this measure. The Standing Committee noted that the validity testing employed a circular comparison to a composite that includes this measure as a component. A Standing Committee member suggested that the developer use the logic model provided in the evidence section as a validation tool for the measure. The developer appreciated the feedback but shared that it is difficult to find comparison measures and to get data to validate processes. They further noted that processes do not always fully correlate with outcomes. The developer shared that they have recently gained access to the results of patient-reported outcome performance measures (PRO-PMs) related to THA/TKA and are working to analyze the relationship with this measure.

The discussion then turned to the risk model. The Standing Committee noted that the c-statistic of 0.65 indicates a poor fit, to which the developer responded that this result indicates that outcomes on this measure are more reflective of quality of care delivered by the facility and not strongly related to patient factors. The Standing Committee noted that both the SMP and public commenters had raised

questions regarding the lack of risk adjustment for social risk factors, noting that the odds ratios for some social factors are larger than those for some clinical factors. Given the elective nature of THA/TKA procedures, the Standing Committee was concerned that patient selection could result in increased disparities and access issues if social risk is not adequately addressed in the risk adjustment. The developer provided additional information on their approach to risk model development, stating that they look at patient-level clinical variables first and then social risk factors. They shared that when the impact of social risk factors is examined in a multivariate model (as opposed to individually), the odds ratios decrease significantly. They further shared that when considering risk factors to include, they consider which factors a hospital can influence. They shared that hospitals participating in the Comprehensive Care for Joint Replacement model through the CMS Innovation Center have demonstrated that hospitals are able to effectively address issues related to social risk. The developer noted that hospital results are highly correlated both with and without the risk factor adjustment. These considerations, coupled with the report from the Office of the Assistant Secretary for Planning and Evaluation (ASPE) of the U.S. Department of Health and Human Services (HHS) advising against adjustment for social risk factors for public reporting, led to the decision not to include social risk factors in the risk adjustment model.

The Standing Committee expressed no concerns with the feasibility or use and usability of the measure. Discussion of related measures was deferred to the post-comment web meeting.

NQF #1551 Hospital-Level 30-Day Risk-Standardized Readmission Rate (RSRR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Yale CORE)/CMS)

Description: The measure estimates a hospital-level risk-standardized readmission rate (RSRR) following elective primary THA and/or TKA in Medicare Fee-For-Service (FFS) 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.; **Measure Type**: Outcome; **Level of Analysis**: Facility; **Setting of Care**: Inpatient/Hospital; **Data Source**: Claims, Enrollment Data

Measure Steward/Developer Representatives at the Meeting

Duwa Amin, MPH – Yale CORE
Darinka Djordjevic, PhD – Yale CORE
Kashika Sahay, PhD, MPH – Yale CORE
Anna Sigler, MPH – Yale CORE
Huihui Yu, PhD – Yale CORE
Sapha Hassan, MPH – Yale CORE
Kristina Gaffney, BS – Yale CORE
Lisa Suter, MD – Yale CORE
Elizabeth Triche, PhD – Yale CORE
Doris Peter, PhD – Yale CORE
James Poyer, MS, MBA – CMS

Standing Committee Votes

- Evidence: Pass-17; No Pass-0 (denominator = 17)
- Performance Gap: H-1; M-16; L-0; I-0 (denominator = 17)
- This measure is deemed as complex and Scientific Acceptability was evaluated by the SMP. The SMP rated the measure as moderate for both reliability and validity.
    - Because voting was conducted after the meeting using an online voting tool, the

Standing Committee voted on the criteria rather than on whether to accept the SMP's ratings.

  - The Standing Committee's rating for Reliability: H-1; M-15; L-0; I-0 (denominator = 16, due to SMP member recusal)
  - The Standing Committee's rating for Validity: H-0; M-15; L-1; I-0 (denominator = 16, due to SMP member recusal)

- Feasibility: H-3; M-13; L-1; I-0 (denominator = 17)
- Use: Pass-17; No Pass-0 (denominator = 17)
- Usability: H-0; M-17; L-0; I-0 (denominator = 17)

Standing Committee Recommendation for Endorsement: Yes-17; No-0 (Denominator = 17)

The Standing Committee recommended the measure for continued endorsement.

This measure was discussed during the second measure evaluation web meeting. Since quorum was not met during the meeting, the Standing Committee discussed all criteria and then voted after the meeting using an online voting tool. This measure estimates a hospital-level, risk-standardized readmission rate (RSRR) following elective primary THA and/or TKA in Medicare FFS beneficiaries who are 65 years of age and older. Originally endorsed in 2012, and most recently endorsed in 2017, the outcome (readmission) is defined as an unplanned readmission for any cause within 30 days of the discharge date for the index admission (i.e., the admission included in the measure cohort).

The Standing Committee noted the evidence was directionally the same yet stronger than the evidence from the previous submission. The Standing Committee questioned whether the performance gap was sufficient to justify continued active endorsement, with 98 percent of facilities performing no different than expected. The developer shared that CMS has criteria for the removal of topped out measures from its programs and that this measure does not meet CMS' criteria for being topped out.

The Standing Committee noted that NQF #1551 received similar public comments to those for NQF #1550 and that the reliability discussion for NQF #1550 also applies to this measure. A Standing Committee member questioned whether the measure could be expanded to even lower-volume hospitals to provide feedback on their performance. A CMS representative clarified that all hospitals are included in the measure calculations and receive feedback reports from CMS. They shared that CMS' goal is to assess as many hospitals as possible but that at very small numbers, one event influences the results, making it difficult to interpret results reliably.

The Standing Committee noted that the entire validity discussion for NQF #1550, including the discussion of the risk model, applies to NQF #1551 as well. The developer shared that for readmissions measures, such as this one, U.S. Congress has mandated that results be stratified into five categories by dual-eligible status.

The Standing Committee expressed no concerns with the feasibility or use and usability of the measure. A Standing Committee member suggested providing context for the measure when it is publicly reported to help patients understand the impact and implication of a readmission. They also felt a low-volume indicator could be useful for the context of results. Discussion of related measures was deferred to the post-comment web meeting.

Public Comment

No public or NQF member comments were provided during the measure evaluation meeting.

Next Steps

NQF will post the draft technical report on April 1, 2021, for public comment for 30 calendar days. The continuous public commenting period with member support will close on April 30, 2021. NQF will reconvene the Standing Committee for the post-comment web meeting on June 1, 2021.

Pre-Evaluation Comments

Comments received as of January 26, 2021.

Topic	Commenter	Comment
NQF #0117 Beta Blockade at Discharge	The Society of Thoracic Surgeons	STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: Definitions for low- and high-performance groups The preliminary analyses for these three process measures found that "It is unclear how low and high-performance groups were defined" for known-group validity testing. This is in reference to the "low performance." "mid performance," and "high performance" categories to which we refer in sect. 2b1.3 in the testing forms. The definitions of these categories are as described in sect. 2b4.1: "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 [confidence interval] 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 high-, low-, and mid-performance groups are thus comparable to the STS "star rating" categories ("higher-than-expected," "lower-than-expected," "as-expected"), although the star ratings are applied to STS composite (outcome) measures only, not to individual process measures. STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: "Insufficient" ratings for Validity We are aware that the NQF validity evaluation algorithm calls for other analyses (sensitivity, specificity, positive predictive value, negative predictive value) in addition to percent agreement. We believe, however, that the validity of our measures at the data element level is adequately demonstrated by the results of the exceptional external audit process that the STS has conducted annually since 2006. The STS audit of the Adult Cardiac Surgery Database (ACSD) is designed to evaluate the accuracy, consistency and comprehensiveness of data collection, and ultimately validate the integrity of the data stored in the Database. Each year, 10% of active ACSD participant sites are randomly selected for audit. In order to evaluate the comprehensiveness of the Database, a list of all cases that are submitted to our analytics center (Duke Clinical Research Institute [DCRI]) from three randomly selected months are compared to the hospital logs of all cases that are performed that year. The data managers provide the auditors with documentation of all cases performed. Each site must demonstrate an effective process to assure that all eligible cases are submitted to the Database.

Topic	Commenter	Comment
		DCRI randomly selects 20 CABG-only and 10 isolated valve cases that are performed in the calendar year for audit at each site; 12 CABG-only and 8 isolated valve cases are re-abstracted at each site. An over-sample is provided to allow for the possibility that a medical record cannot be located by the site and is therefore unavailable for re-abstraction. A specified group of data variables are evaluated each year, utilizing the current version of the STS Adult Cardiac Surgery Data Specifications; the number of variables increases every year. (For example, 82 variables were evaluated in 2015; 86 in 2017; 91 in 2019.) Agreement rates are calculated for each of the individual variables, each variable category and overall. The overall aggregate agreement rate for the most recent five audits is shown in the table below: Audit Year Total Cases Total Mismatch Overall Aggregate Agreement Rate 2019 203,840 14,313 92.98% 2018 222,500 10,346 95.35% 2017 144,920 5,010 96.54% 2016 144,368 5,494 96.19% 2015 141,047 5,409 96.17% These results, and the rigorous audit process through which they are obtained, demonstrate the accuracy and completeness of the data in the STS ACSD. This conclusion is further supported by comments received from our external auditors in each year's final audit report. Two examples follow: [2015] "There were 141,047 total variables abstracted and there were 135,638 variables that matched, resulting in an overall agreement rate of 96.17% (95.73% in 2014). This overall performance rate reflects a high level of accuracy in data collection and evidence that the data contained in the ACSD are valid." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Telligen Final Report.Telligen, December 2015. [2018] "The overall aggregate agreement rate was 95.4%, demonstrating that the data contained in the ACSD is both comprehensive and highly accurate… The surgeons and staff that perform the data collection and submission to the ACSD were found to be committed to the STS goal of collecting quality data." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Final Audit Report 2018.Cardiac Registry Support, LLC, November 2019. In summary, we believe that the additional information provided here adequately demonstrates the validity of STS measures 0117, 0127, 0134 at the data element level, and will appreciate a reconsideration of the preliminary "insufficient" rating.

Topic	Commenter	Comment
		STS Response to Preliminary Analyses for Measures 0117 & 0134: "Low" ratings for Opportunity for Improvement We understand but respectfully disagree with the assessment that these two STS measures are "topped out" and therefore subject to loss of endorsement. We ask that you please consider the following: • The STS believes that these evidence based, guideline-directed measures are significantly responsible for the dramatic improvement we have demonstrated in outcomes and in process-of-care compliance, as documented in a 2019 Joint Commission Journal on Quality and Patient Safety article (1). Table 2 shows a 54% improvement in compliance with the Discharge Beta-Blocker measure (#0117) between 2002 and 2016, and a 32% improvement in compliance with the IMA Use measure (#0134) between 1998 and 2016. • It is inappropriate to view these improvements as a rationale to remove endorsement for these measures and risk a deterioration in results due to the perception that these measures are no longer important. Cardiac surgeries are high-stakes procedures in which small errors or deviations from standardized care processes can lead to death. From our perspective, a residual 1-2 % failure rate for individual process measures is not acceptable. • Cardiac surgery is comparable to the airline industry in that we must strive for high reliability; our goal is a 100% success rate. • Even small failure rates may result in a participant rating below the STS average, providing the potential to identify statistically meaningful differences in performance. • Furthermore, the continued use and endorsement of these measures does not contribute to an excessive data entry burden for clinicians or their staff. The data for these processes of care is routinely collected – in a data registry with over 95% participation in the U.S. – for the STS CABG Composite for which these are component measures, along with mortality and morbidity outcomes. Concerns related to measures becoming "topped out" are more relevant to non-registry measures for which data collection may require the allocation of additional resources. We therefore believe that the "topped out" assessment for measures 0117 & 0134 is unwarranted and ask NQF staff and the Surgery Standing Committee to consider a higher Opportunity for Improvement rating for each measure. 1. Shahian DM. Professional Society Leadership in Health Care Quality: The Society of Thoracic Surgeons Experience. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2019;45(7):466-79.
NQF #0127 Preoperative Beta Blockade	The Society of Thoracic Surgeons	STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: Definitions for low- and high-performance groups The preliminary analyses for these three process measures found that "It is unclear how low and high-performance groups were defined" for known-group validity testing. This is in reference to the "low performance." "mid

Topic	Commenter	Comment
		performance," and "high performance" categories to which we refer in sect. 2b1.3 in the testing forms. The definitions of these categories are as described in sect. 2b4.1: "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 [confidence interval] 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 high-, low-, and mid-performance groups are thus comparable to the STS "star rating" categories ("higher-than-expected," "lower-than-expected," "as-expected"), although the star ratings are applied to STS composite (outcome) measures only, not to individual process measures. STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: "Insufficient" ratings for Validity We are aware that the NQF validity evaluation algorithm calls for other analyses (sensitivity, specificity, positive predictive value, negative predictive value) in addition to percent agreement. We believe, however, that the validity of our measures at the data element level is adequately demonstrated by the results of the exceptional external audit process that the STS has conducted annually since 2006. The STS audit of the Adult Cardiac Surgery Database (ACSD) is designed to evaluate the accuracy, consistency and comprehensiveness of data collection, and ultimately validate the integrity of the data stored in the Database. Each year, 10% of active ACSD participant sites are randomly selected for audit. In order to evaluate the comprehensiveness of the Database, a list of all cases that are submitted to our analytics center (Duke Clinical Research Institute [DCRI]) from three randomly selected months are compared to the hospital logs of all cases that are performed that year. The data managers provide the auditors with documentation of all cases performed. Each site must demonstrate an effective process to assure that all eligible cases are submitted to the Database. DCRI randomly selects 20 CABG-only and 10 isolated valve cases that are performed in the calendar year for audit at each site; 12 CABG-only and 8 isolated valve cases are re-abstracted at each site. An over-sample is provided to allow for the possibility that a medical record cannot be located by the site and is therefore unavailable for re-abstraction. A specified group of data variables are evaluated each year, utilizing the current version of the STS Adult Cardiac Surgery Data Specifications; the number of variables increases every year. (For example, 82 variables were evaluated in 2015; 86 in 2017; 91 in 2019.) Agreement rates are calculated for each of the individual variables, each variable category and overall. The overall aggregate agreement rate for the most recent five audits is shown in the table below: Audit Year Total Cases Total Mismatch Overall Aggregate Agreement Rate

Topic	Commenter	Comment
		2019 203,840 14,313 92.98% 2018 222,500 10,346 95.35% 2017 144,920 5,010 96.54% 2016 144,368 5,494 96.19% 2015 141,047 5,409 96.17% These results, and the rigorous audit process through which they are obtained, demonstrate the accuracy and completeness of the data in the STS ACSD. This conclusion is further supported by comments received from our external auditors in each year's final audit report. Two examples follow: [2015] "There were 141,047 total variables abstracted and there were 135,638 variables that matched, resulting in an overall agreement rate of 96.17% (95.73% in 2014). This overall performance rate reflects a high level of accuracy in data collection and evidence that the data contained in the ACSD are valid." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Telligen Final Report.Telligen, December 2015. [2018] "The overall aggregate agreement rate was 95.4%, demonstrating that the data contained in the ACSD is both comprehensive and highly accurate… The surgeons and staff that perform the data collection and submission to the ACSD were found to be committed to the STS goal of collecting quality data." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Final Audit Report 2018.Cardiac Registry Support, LLC, November 2019. In summary, we believe that the additional information provided here adequately demonstrates the validity of STS measures 0117, 0127, 0134 at the data element level, and will appreciate a reconsideration of the preliminary "insufficient" rating.
NQF #0134 Use of Internal Mammary Artery (IMA) in Coronary Artery Bypass Graft (CABG)	The Society of Thoracic Surgeons	STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: Definitions for low- and high-performance groups The preliminary analyses for these three process measures found that "It is unclear how low and high-performance groups were defined" for known-group validity testing. This is in reference to the "low performance." "mid performance," and "high performance" categories to which we refer in sect. 2b1.3 in the testing forms. The definitions of these categories are as described in sect. 2b4.1: "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 [confidence interval] lies entirely above/below the STS average. The

Topic	Commenter	Comment
		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 high-, low-, and mid-performance groups are thus comparable to the STS "star rating" categories ("higher-than-expected," "lower-than-expected," "as-expected"), although the star ratings are applied to STS composite (outcome) measures only, not to individual process measures. STS Response to Preliminary Analyses for Measures 0117, 0127, 0134: "Insufficient" ratings for Validity We are aware that the NQF validity evaluation algorithm calls for other analyses (sensitivity, specificity, positive predictive value, negative predictive value) in addition to percent agreement. We believe, however, that the validity of our measures at the data element level is adequately demonstrated by the results of the exceptional external audit process that the STS has conducted annually since 2006. The STS audit of the Adult Cardiac Surgery Database (ACSD) is designed to evaluate the accuracy, consistency and comprehensiveness of data collection, and ultimately validate the integrity of the data stored in the Database. Each year, 10% of active ACSD participant sites are randomly selected for audit. In order to evaluate the comprehensiveness of the Database, a list of all cases that are submitted to our analytics center (Duke Clinical Research Institute [DCRI]) from three randomly selected months are compared to the hospital logs of all cases that are performed that year. The data managers provide the auditors with documentation of all cases performed. Each site must demonstrate an effective process to assure that all eligible cases are submitted to the Database. DCRI randomly selects 20 CABG-only and 10 isolated valve cases that are performed in the calendar year for audit at each site; 12 CABG-only and 8 isolated valve cases are re-abstracted at each site. An over-sample is provided to allow for the possibility that a medical record cannot be located by the site and is therefore unavailable for re-abstraction. A specified group of data variables are evaluated each year, utilizing the current version of the STS Adult Cardiac Surgery Data Specifications; the number of variables increases every year. (For example, 82 variables were evaluated in 2015; 86 in 2017; 91 in 2019.) Agreement rates are calculated for each of the individual variables, each variable category and overall. The overall aggregate agreement rate for the most recent five audits is shown in the table below: Audit Year Total Cases Total Mismatch Overall Aggregate Agreement Rate 2019 203,840 14,313 92.98% 2018 222,500 10,346 95.35% 2017 144,920 5,010 96.54% 2016 144,368 5,494 96.19%

Topic	Commenter	Comment
		2015 141,047 5,409 96.17% These results, and the rigorous audit process through which they are obtained, demonstrate the accuracy and completeness of the data in the STS ACSD. This conclusion is further supported by comments received from our external auditors in each year's final audit report. Two examples follow: [2015] "There were 141,047 total variables abstracted and there were 135,638 variables that matched, resulting in an overall agreement rate of 96.17% (95.73% in 2014). This overall performance rate reflects a high level of accuracy in data collection and evidence that the data contained in the ACSD are valid." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Telligen Final Report.Telligen, December 2015. [2018] "The overall aggregate agreement rate was 95.4%, demonstrating that the data contained in the ACSD is both comprehensive and highly accurate… The surgeons and staff that perform the data collection and submission to the ACSD were found to be committed to the STS goal of collecting quality data." Source: The Society of Thoracic Surgeons Adult Cardiac Surgery Database Audit – Final Audit Report 2018.Cardiac Registry Support, LLC, November 2019. In summary, we believe that the additional information provided here adequately demonstrates the validity of STS measures 0117, 0127, 0134 at the data element level, and will appreciate a reconsideration of the preliminary "insufficient" rating. STS Response to Preliminary Analyses for Measures 0117 & 0134: "Low" ratings for Opportunity for Improvement We understand but respectfully disagree with the assessment that these two STS measures are "topped out" and therefore subject to loss of endorsement. We ask that you please consider the following: • The STS believes that these evidence based, guideline-directed measures are significantly responsible for the dramatic improvement we have demonstrated in outcomes and in process-of-care compliance, as documented in a 2019 Joint Commission Journal on Quality and Patient Safety article (1). Table 2 shows a 54% improvement in compliance with the Discharge Beta-Blocker measure (#0117) between 2002 and 2016, and a 32% improvement in compliance with the IMA Use measure (#0134) between 1998 and 2016. • It is inappropriate to view these improvements as a rationale to remove endorsement for these measures and risk a deterioration in results due to the perception that these measures are no longer important. Cardiac surgeries are high-

Topic	Commenter	Comment
		stakes procedures in which small errors or deviations from standardized care processes can lead to death. From our perspective, a residual 1-2 % failure rate for individual process measures is not acceptable. • Cardiac surgery is comparable to the airline industry in that we must strive for high reliability; our goal is a 100% success rate. • Even small failure rates may result in a participant rating below the STS average, providing the potential to identify statistically meaningful differences in performance. • Furthermore, the continued use and endorsement of these measures does not contribute to an excessive data entry burden for clinicians or their staff. The data for these processes of care is routinely collected – in a data registry with over 95% participation in the U.S. – for the STS CABG Composite for which these are component measures, along with mortality and morbidity outcomes. Concerns related to measures becoming "topped out" are more relevant to non-registry measures for which data collection may require the allocation of additional resources. We therefore believe that the "topped out" assessment for measures 0117 & 0134 is unwarranted and ask NQF staff and the Surgery Standing Committee to consider a higher Opportunity for Improvement rating for each measure. 1. Shahian DM. Professional Society Leadership in Health Care Quality: The Society of Thoracic Surgeons Experience. Joint Commission journal on quality and patient safety / Joint Commission Resources. 2019;45(7):466-79.
NQF #1550 Hospital-Level Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA)	American Medical Association	The American Medical Association (AMA) appreciates the opportunity to comment on the NQF Measure #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). In reviewing the calculation, we are disappointed to see the minimum measure score reliability result calculated at 0.46 and the intraclass correlation coefficient (ICC) calculated at 0.524 using a minimum case number of just 25 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability and require higher case minimums to allow the overwhelming majority of hospitals to achieve an ICC of 0.6 or higher. The AMA is also extremely concerned that the measure developer used the recommendation to exclude social risk factors in the risk adjustment models for measures that are publicly reported as outlined in the recent report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program (ASPE, 2020). We believe that while the current testing may not have produced results that would indicate incorporation of the two social risk factors included in testing, this measure is currently used both for public reporting and value-based purchasing. A primary limitation of the ASPE report was that none of the recommendations adequately addressed whether it was or appropriate to adjust for social risk factors in the same measure used for more than one accountability purpose, which is the case here. This discrepancy along with the fact

Topic	Commenter	Comment
		that the additional analysis using the American Community Survey is not yet released must be addressed prior to any reliance on the recommendations within this report. In addition, we question whether the measure continues to be useful to distinguish hospital performance and drive improvements based on the distribution of hospital's performance scores where only 60 hospitals performed better than the national rate and 50 hospitals performed worse (as noted in section 2b4 and the discussion on improvement in section 4b1 of the measure submission form), and where there was only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019. We request that the Standing Committee evaluate whether the measure continues to meet the measure evaluation criteria required for endorsement. Reference: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs
NQF #1550 Hospital-Level Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA)	Federation of American Hospitals	The Federation of American Hospitals (FAH) appreciates the opportunity to comment on Measure #1550, Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). The FAH is concerned that even though the median reliability score was 0.87 for hospitals with at least 25 cases, reliability ranged from 0.46 to 1.00 and that the intraclass correlation coefficients (ICC) was 0.524. The FAH believes that the developer must increase the minimum sample size to a higher number to produce a minimum reliability threshold of sufficient magnitude (e.g. 0.7 or higher) and an ICC of 0.6 or higher. In addition, the FAH is very concerned to see that the measure developer's rationale to not include social risk factors in the risk adjustment model was in part based on the recommendations from the report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program released in March of last year (ASPE, 2020). A fundament flaw within the ASPE report was the lack of any recommendation addressing how a single measure with multiple accountability uses should address inclusion of social risk factors as is the case with this measure, which is both publicly reported and included in the Hospital Value-Based Purchasing program. Regardless of whether the testing of social risk factors produced results that were sufficiently significant, the FAH believes that no developer should rely on the recommendations of this report until the question of how to handle multiple uses is addressed along with the additional analysis using the American Community Survey.

Topic	Commenter	Comment
		Lastly, the FAH is concerned that there is insufficient variation in performance across hospitals and limited opportunities for improvement to support this measure's continued use in accountability programs. Specifically, the performance scores reported in 2b4. Identification of Statistically Significant and Meaningful Difference in Performance are generally low with only 60 hospitals identified as better than the national rate and 50 are worse than the national rate. We base our concerns on these results along with the discussion on improvement in section 4b1 of the measure submission form where only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019 was found. As a result, the FAH requests that the Standing Committee carefully consider whether the measure as specified should continue to be endorsed. Reference: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs
NQF #1551 Hospital-Level 30-Day Risk-Standardized Readmission Rate (RSRR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA)	American Medical Association	The American Medical Association (AMA) appreciates the opportunity to comment on NQF Measure #1551, Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). We are disappointed to see the minimum measure score reliability results calculated at 0.29 and the intraclass correlation coefficient (ICC) calculated at 0.454 using a minimum case number of just 25 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability and require higher case minimums to allow the overwhelming majority of hospitals to achieve an ICC of 0.6 or higher. In reviewing the calculation, the AMA is also extremely concerned to see that the measure developer used the recommendation to not include social risk factors in the risk adjustment models for measures that are publicly reported as outlined in the recent report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program (ASPE, 2020). We believe that while the current testing may not have produced results that would indicate incorporation of the two social risk factors included in testing, this measure is currently used both for public reporting and value-based purchasing. A primary limitation of the ASPE report was that none of the recommendations adequately addressed whether it was appropriate to adjust for social risk factors in the same measure used for more than one accountability purpose, which is the case here. This discrepancy along with the fact that the additional analysis using the American Community Survey is not yet released must be addressed prior to any reliance on the recommendations within this report.

Topic	Commenter	Comment
		In addition, we question whether the measure continues to be useful to distinguish hospital performance and drive improvements based on the distribution of hospital's performance scores where only 44 hospitals performed better than the national rate and 24 hospitals were worse (as noted in section 2b4 and the discussion on improvement in section 4b1 of the measure submission form), and where there was only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019. We request that the Standing Committee evaluate whether the measure continues to meet the measure evaluation criteria required for endorsement. Reference: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs
NQF #1551 Hospital-Level 30-Day Risk-Standardized Readmission Rate (RSRR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA)	Federation of American Hospitals	The Federation of American Hospitals (FAH) appreciates the opportunity to comment on Measure #1551, Hospital-level risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). The FAH is concerned that even though the median reliability score was 0.77 for hospitals with at least 25 cases, reliability ranged from 0.29 to 0.99 and that the intraclass correlation coefficients (ICC) was 0.454. The FAH believes that the developer must increase the minimum sample size to a higher number to produce a minimum reliability threshold of sufficient magnitude (e.g. 0.7 or higher) and an ICC of 0.6 or higher. In addition, the FAH is very concerned to see that the measure developer's rationale to not include social risk factors in the risk adjustment model was in part based on the recommendations from the report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program released in March of last year (ASPE, 2020). A fundament flaw within the ASPE report was the lack of any recommendation addressing how a single measure with multiple accountability uses should address inclusion of social risk factors as is the case with this measure, which is both publicly reported and included in the Hospital Value-Based Purchasing program. Regardless of whether the testing of social risk factors produced results that were sufficiently significant, the FAH believes that no developer should rely on the recommendations of this report until the question of how to handle multiple uses is addressed along with the additional analysis using the American Community Survey. Lastly, the FAH is concerned that there is insufficient variation in performance across hospitals and limited opportunities for improvement to support this measure's continued use in accountability programs. Specifically, the performance scores reported in 2b4. Identification of Statistically Significant and Meaningful Difference in Performance are generally low with only 44 hospitals identified as better than the national rate and 24 are worse than the national rate. We base

<table>
<tr><th>Topic</th><th>Commenter</th><th>Comment</th></tr>
<tr><td></td><td></td><td>our concerns on these results along with the discussion on improvement in section 4b1 of the measure submission form where only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019 was found.
As a result, the FAH requests that the Standing Committee carefully consider whether the measure as specified should continue to be endorsed.

Reference:
Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare’s Value-Based Purchasing Program.2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs</td></tr>
<tr><td>NQF #3030 STS Individual Surgeon Composite Measure for Adult Cardiac Surgery</td><td>The Society of Thoracic Surgeons</td><td>STS Updates to Measure Testing Document Section 1b.4
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.
In order to shed light on disparities, we used logistic regression to study the associations of race, ethnicity and insurance status with operative mortality and major morbidity while adjusting for covariates included in any of the 2018 risk adjustment models (see other sections for details of covariate adjustment – we used the most recent 2018 CABG, valve and valve+CABG models for mortality and major morbidity). Odds ratios with 95% confidence intervals (CI's) and p-values are summarized in the table below.
<table>
<tr><th></th><th>Mortality</th><th></th><th>Major Morbidity</th><th></th></tr>
<tr><th></th><th>Adjusted OR (95% CI)</th><th>p-value</th><th>Adjusted OR (95% CI)</th><th>p-value</th></tr>
<tr><td>Insurance status among patients age>=65</td><td></td><td></td><td></td><td></td></tr>
<tr><td>Medicare without Medicaid/Commercial-HMO</td><td>(ref)</td><td></td><td>(ref)</td><td></td></tr>
<tr><td>Medicare Medicaid dual eligible</td><td>0.95 (0.87, 1.03)</td><td>0.2178</td><td>1.05 (1.00, 1.09)</td><td>0.0537</td></tr>
<tr><td>Medicare Commercial-HMO without Medicaid</td><td>0.93 (0.89, 0.97)</td><td>0.0003</td><td>0.97 (0.95, 0.99)</td><td>0.0095</td></tr>
<tr><td>Commercial-HMO without Medicare</td><td>0.97 (0.90, 1.05)</td><td>0.448</td><td>1.00 (0.96, 1.04)</td><td>0.9403</td></tr>
</table></td></tr>
</table>

Topic	Commenter	Comment

Insurance status among patients age<65

Commercial-HMO without Medicare/Medicaid	(ref)		(ref)	
Medicare or Medicaid	1.08 (1.01, 1.17)	0.0332	1.16 (1.12, 1.19)	<.0001
None/Self Paid	1.10 (0.98, 1.22)	0.099	1.08 (1.03, 1.13)	0.0022
Other	1.11 (0.96, 1.28)	0.151	1.03 (0.96, 1.09)	0.4283
Black Race	1.01 (0.95, 1.07)	0.8042	1.18 (1.15, 1.22)	<.0001
Hispanic ethnicity	1.00 (0.94, 1.07)	0.9194	1.01 (0.97, 1.04)	0.6444

STS Response to Preliminary Analyses for Measures 3030, 3031, 3032: “Insufficient” ratings for Validity

For each of these composite measures, the Preliminary Analysis states that “Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity.” As in past endorsement and endorsement maintenance reviews for our composite measures, we believe it to be a reasonable approach to use our morbidity and mortality component scores as the “gold standard” against which to demonstrate construct or criterion validity of the composite scores across our three performance categories: “higher-than-expected,” “lower-than-expected,” and “as-expected” (as defined in 2b1.2 in our composite testing forms). If participants/surgeons with “higher-than-expected” composite ratings have consistently lower risk-adjusted mortality and lower risk-adjusted morbidity compared to participants/surgeons with “lower-than-expected” ratings, we believe the validity of the composite score is demonstrated. The STS has the most sophisticated outcomes data and methodology available for heart surgery, in a database with over 95% penetration across cardiac surgery practices in the U.S.; we therefore have no other “gold standard” against which to compare our results.

NQF staff have suggested the use of an external standard – e.g., a measure for a different cardiothoracic surgery procedure – for testing the validity of our composite measures. However, published studies have shown that excellent performance on one surgical procedure does not necessarily correlate with excellent performance on another procedure. We therefore maintain that the approach described above is appropriate for demonstrating the validity of our composite measures.

STS Updates to Measure Testing Document Section 1.8 for Measures NQF#s 3030, 3031, 3032 - PART 1

Topic	Commenter	Comment
		1.8 What were the social risk factors that were available and analyzed? The STS position on inclusion of social risk factors (e.g., SES/SDS/race) as risk model variables is best summarized in this excerpt from our 2018 risk model publication [1]. We describe in detail the controversies about such variables, and how we have attempted to reconcile them: "Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (eg, readmission) than for others (hospital mortality, complications). Notably, as part of a National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure and found minimal impact. In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator [Note: nor as a surrogate for such factors]. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (eg, genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission)." STS is aware of the recent NEJM paper by Vyas and colleagues [2] and has directly communicated with the lead author to explain why race is included in STS models, and to correct several misinterpretations and misrepresentations in this article. Dr. Vyas acknowledged that they included extended quotes from our risk model paper precisely because we were one of the few risk model developers that thoroughly described our rationale for race inclusion, as noted in the excerpt above. Documents produced by NQF [3, 4], the National Academy of Medicine [5-8], the Office of the Assistant Secretary for Planning and Evaluation (Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs) [9], and as part of the 21st Century Cures Act legislation [10] are particularly instructive. They summarize the arguments for

Topic	Commenter	Comment
		and against inclusion of SDS/SES/racial adjustment in risk models; context-specific considerations for when they might be appropriate or inappropriate; strategies to avoid the potential adverse unintended consequences of such adjustment; concomitant monitoring for social and racial inequities through stratification; and special approaches for providers who care for high proportions of disadvantaged populations (e.g., payment adjustments, additional resources). Adjustment for SDS/SES/racial factors has generally been regarded as acceptable (e.g., in NQF white papers) when there is both an empirical association AND a plausible conceptual association of the risk variable with an outcome. For example, an SES/SDS/racial risk factor might be appropriate as a risk variable for readmission or mortality risk models, but not for CAUTI (catheter-associated urinary tract infections), CLABSI (central line-associated bloodstream infection), or process measures. For many outcomes, SES/SDS/racial adjustment is warranted to optimize risk model accuracy. For example, recent STS and Duke Clinical Research Institute analyses show that if race variables are excluded from some STS models, the resulting outcomes estimates are markedly different than the actual observed outcomes, and the O/E ratios are significantly different than unity, especially when the models are applied to racial minority subpopulations—in other words, the models are less well calibrated, an essential feature of any risk model. This miscalibration persisted even when an SES/SDS indicator (specifically, dual eligible status) was simultaneously included in the models (i.e., thus addressing the hypothesis that the putative association of race and various outcomes is actually mediated by SES/SDS). Use of risk estimates from such models for patient counseling and shared decision-making would be misleading to patients and would inaccurately portray (and unfairly disadvantage) the risk-adjusted performance of providers, especially those caring for minority populations. Importantly, STS and its analytic center re-estimate risk factor coefficients several times annually, so that any changes in the association of race with outcomes will be implemented in the newest estimates. Further, STS is geocoding it adult cardiac surgery records and will use this information to derive an Area Deprivation Index for all patients with a valid address, thus providing us with the ability to further study the impact of race and SES/SDS using the what is arguably the most sensitive and comprehensive SES/SDS indicator. Finally, STS is aware of the recommendation in the ASPE report of October 2020 that functional status indicators be included in risk models as it may account for some of the impact on outcomes associated that is currently attributed to race. Although STS has a well-documented frailty indicator (5 meter walk test), it has not been collected with sufficient consistency by our participants to allow its inclusion in our models. Accordingly, STS has established a new working group on Frailty/functional indicators whose goal is to develop a new indicator that can be captured for virtually all patients using a combination of history, lab data, functional status, etc. Once developed, it will be added to STS models. Although SDS/SES/racial risk adjustment may be indicated to assure optimal risk model estimates based on current data, it is widely believed that such adjustment could potentially obscure disparities in care. To avoid this potential unintended consequence, most of the national guidance documents cited above recommend that any risk model results

Topic	Commenter	Comment
		that are adjusted for SES/SDS/racial factors also present concomitant results in which outcomes are stratified by the same variables. This is a much more direct and explicit approach to monitor disparities and inequities and has been followed by STS in its risk modeling and performance measures. Please refer to the race-specific disparities data provided for each of the domains (mortality and morbidity) of measure 3030 under question 1b.4 (Importance tab) of the submission form (to be completed by the November submission deadline), which we believe will suffice to comply with this recommendation. STS Updates to Measure Testing Document Section 1.8 for Measures NQF #s 3030, 3031, 3032 - PART 2 1.8 What were the social risk factors that were available and analyzed? 1. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411-8. 2. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020. 3. National Quality Forum. Risk adjustment for Socioeconomic Status or other Sociodemographic Factors, accessed at http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx on June 24, 2020. 2014. 4. The National Quality Forum. Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors. January 15, 2017. Available from: https://www.qualityforum.org/Publications/2017/07/Social_Risk_Trial_Final_Report.aspx. 5. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment. Washington, DC: The National Academies Press; 2017. 6. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment: Data. Washington, DC; 2016. 7. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: The National Academies Press; 2016. 8. National Academies of Sciences, Engineering, Medicine,. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington, DC: The National Academies Press; 2016. 110 p.

Topic	Commenter	Comment
		9. Office of the Assistant Secretary for Planning and Evaluation USDoHaHS. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014. Washington, DC; 2016. 10. 114th Congress of the United States. 21st Century Cures Act (Public Law 114–255). Washington, DC; 2016.
STS Mitral Valve Repair/Replacement (MVRR) Composite Score	The Society of Thoracic Surgeons	STS Response to Preliminary Analyses for Measures 3030, 3031, 3032: "Insufficient" ratings for Validity For each of these composite measures, the Preliminary Analysis states that "Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity." As in past endorsement and endorsement maintenance reviews for our composite measures, we believe it to be a reasonable approach to use our morbidity and mortality component scores as the "gold standard" against which to demonstrate construct or criterion validity of the composite scores across our three performance categories: "higher-than-expected," "lower-than-expected," and "as-expected" (as defined in 2b1.2 in our composite testing forms). If participants/surgeons with "higher-than-expected" composite ratings have consistently lower risk-adjusted mortality and lower risk-adjusted morbidity compared to participants/surgeons with "lower-than-expected" ratings, we believe the validity of the composite score is demonstrated. The STS has the most sophisticated outcomes data and methodology available for heart surgery, in a database with over 95% penetration across cardiac surgery practices in the U.S.; we therefore have no other "gold standard" against which to compare our results. NQF staff have suggested the use of an external standard – e.g., a measure for a different cardiothoracic surgery procedure – for testing the validity of our composite measures. However, published studies have shown that excellent performance on one surgical procedure does not necessarily correlate with excellent performance on another procedure. We therefore maintain that the approach described above is appropriate for demonstrating the validity of our composite measures. STS Updates to Measure Testing Document Section 1.8 for Measures NQF#s 3030, 3031, 3032 - PART 1 1.8 What were the social risk factors that were available and analyzed? The STS position on inclusion of social risk factors (e.g., SES/SDS/race) as risk model variables is best summarized in this excerpt from our 2018 risk model publication [1]. We describe in detail the controversies about such variables, and how we have attempted to reconcile them: "Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic factors (eg, race, ethnicity, education, income, payer [eg, Medicare-

Topic	Commenter	Comment
		Medicaid dual eligible status]) is a topic of intense health policy debate. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (eg, readmission) than for others (hospital mortality, complications). Notably, as part of a National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure and found minimal impact. In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator [Note: nor as a surrogate for such factors]. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (eg, genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission)." STS is aware of the recent NEJM paper by Vyas and colleagues [2] and has directly communicated with the lead author to explain why race is included in STS models, and to correct several misinterpretations and misrepresentations in this article. Dr. Vyas acknowledged that they included extended quotes from our risk model paper precisely because we were one of the few risk model developers that thoroughly described our rationale for race inclusion, as noted in the excerpt above. Documents produced by NQF [3, 4], the National Academy of Medicine [5-8], the Office of the Assistant Secretary for Planning and Evaluation (Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs) [9], and as part of the 21st Century Cures Act legislation [10] are particularly instructive. They summarize the arguments for and against inclusion of SDS/SES/racial adjustment in risk models; context-specific considerations for when they might be appropriate or inappropriate; strategies to avoid the potential adverse unintended consequences of such adjustment; concomitant monitoring for social and racial inequities through stratification; and special approaches for providers who care for high proportions of disadvantaged populations (e.g., payment adjustments, additional resources). Adjustment for SDS/SES/racial factors has generally been regarded as acceptable (e.g., in NQF white papers) when there is both an empirical association AND a plausible conceptual association of the risk variable with an outcome. For example, an SES/SDS/racial risk factor might be appropriate as a risk variable for readmission or mortality risk models,

Topic	Commenter	Comment
		but not for CAUTI (catheter-associated urinary tract infections), CLABSI (central line-associated bloodstream infection), or process measures. For many outcomes, SES/SDS/racial adjustment is warranted to optimize risk model accuracy. For example, recent STS and Duke Clinical Research Institute analyses show that if race variables are excluded from some STS models, the resulting outcomes estimates are markedly different than the actual observed outcomes, and the O/E ratios are significantly different than unity, especially when the models are applied to racial minority subpopulations—in other words, the models are less well calibrated, an essential feature of any risk model. This miscalibration persisted even when an SES/SDS indicator (specifically, dual eligible status) was simultaneously included in the models (i.e., thus addressing the hypothesis that the putative association of race and various outcomes is actually mediated by SES/SDS). Use of risk estimates from such models for patient counseling and shared decision-making would be misleading to patients and would inaccurately portray (and unfairly disadvantage) the risk-adjusted performance of providers, especially those caring for minority populations. Importantly, STS and its analytic center re-estimate risk factor coefficients several times annually, so that any changes in the association of race with outcomes will be implemented in the newest estimates. Further, STS is geocoding it adult cardiac surgery records and will use this information to derive an Area Deprivation Index for all patients with a valid address, thus providing us with the ability to further study the impact of race and SES/SDS using the what is arguably the most sensitive and comprehensive SES/SDS indicator. Finally, STS is aware of the recommendation in the ASPE report of October 2020 that functional status indicators be included in risk models as it may account for some of the impact on outcomes associated that is currently attributed to race. Although STS has a well-documented frailty indicator (5 meter walk test), it has not been collected with sufficient consistency by our participants to allow its inclusion in our models. Accordingly, STS has established a new working group on Frailty/functional indicators whose goal is to develop a new indicator that can be captured for virtually all patients using a combination of history, lab data, functional status, etc. Once developed, it will be added to STS models. Although SDS/SES/racial risk adjustment may be indicated to assure optimal risk model estimates based on current data, it is widely believed that such adjustment could potentially obscure disparities in care. To avoid this potential unintended consequence, most of the national guidance documents cited above recommend that any risk model results that are adjusted for SES/SDS/racial factors also present concomitant results in which outcomes are stratified by the same variables. This is a much more direct and explicit approach to monitor disparities and inequities and has been followed by STS in its risk modeling and performance measures. Please refer to the race-specific disparities data provided for each of the domains (mortality and morbidity) of measure 3030 under question 1b.4 (Importance tab) of the submission form (to be completed by the November submission deadline), which we believe will suffice to comply with this recommendation.

Topic	Commenter	Comment
		STS Updates to Measure Testing Document Section 1.8 for Measures NQF #s 3030, 3031, 3032 - PART 2 1.8 What were the social risk factors that were available and analyzed? 1. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411-8. 2. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020. 3. National Quality Forum. Risk adjustment for Socioeconomic Status or other Sociodemographic Factors, accessed at http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx on June 24, 2020. 2014. 4. The National Quality Forum. Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors. January 15, 2017. Available from: https://www.qualityforum.org/Publications/2017/07/Social_Risk_Trial_Final_Report.aspx. 5. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment. Washington, DC: The National Academies Press; 2017. 6. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment: Data. Washington, DC; 2016. 7. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: The National Academies Press; 2016. 8. National Academies of Sciences, Engineering, Medicine,. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington, DC: The National Academies Press; 2016. 110 p. 9. Office of the Assistant Secretary for Planning and Evaluation USDoHaHS. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014. Washington, DC; 2016. 10. 114th Congress of the United States. 21st Century Cures Act (Public Law 114–255). Washington, DC; 2016.
STS Mitral Valve Repair/Replacem	The Society of	STS Response to Preliminary Analyses for Measures 3030, 3031, 3032: "Insufficient" ratings for Validity

Topic	Commenter	Comment
ent (MVRR) + Coronary Artery Bypass Graft (CABG) Composite Score	Thoracic Surgeons	For each of these composite measures, the Preliminary Analysis states that "Demonstrating a relationship between performance on the overall composite and the composite domains may not be a valid assessment of composite score validity." As in past endorsement and endorsement maintenance reviews for our composite measures, we believe it to be a reasonable approach to use our morbidity and mortality component scores as the "gold standard" against which to demonstrate construct or criterion validity of the composite scores across our three performance categories: "higher-than-expected," "lower-than-expected," and "as-expected" (as defined in 2b1.2 in our composite testing forms). If participants/surgeons with "higher-than-expected" composite ratings have consistently lower risk-adjusted mortality and lower risk-adjusted morbidity compared to participants/surgeons with "lower-than-expected" ratings, we believe the validity of the composite score is demonstrated. The STS has the most sophisticated outcomes data and methodology available for heart surgery, in a database with over 95% penetration across cardiac surgery practices in the U.S.; we therefore have no other "gold standard" against which to compare our results. NQF staff have suggested the use of an external standard – e.g., a measure for a different cardiothoracic surgery procedure – for testing the validity of our composite measures. However, published studies have shown that excellent performance on one surgical procedure does not necessarily correlate with excellent performance on another procedure. We therefore maintain that the approach described above is appropriate for demonstrating the validity of our composite measures. STS Updates to Measure Testing Document Section 1.8 for Measures NQF#s 3030, 3031, 3032 - PART 1 1.8 What were the social risk factors that were available and analyzed? The STS position on inclusion of social risk factors (e.g., SES/SDS/race) as risk model variables is best summarized in this excerpt from our 2018 risk model publication [1]. We describe in detail the controversies about such variables, and how we have attempted to reconcile them: "Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (eg, readmission) than for others (hospital mortality, complications). Notably, as

Topic	Commenter	Comment
		part of a National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure and found minimal impact. In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator [Note: nor as a surrogate for such factors]. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (eg, genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission)." STS is aware of the recent NEJM paper by Vyas and colleagues [2] and has directly communicated with the lead author to explain why race is included in STS models, and to correct several misinterpretations and misrepresentations in this article. Dr. Vyas acknowledged that they included extended quotes from our risk model paper precisely because we were one of the few risk model developers that thoroughly described our rationale for race inclusion, as noted in the excerpt above. Documents produced by NQF [3, 4], the National Academy of Medicine [5-8], the Office of the Assistant Secretary for Planning and Evaluation (Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs) [9], and as part of the 21st Century Cures Act legislation [10] are particularly instructive. They summarize the arguments for and against inclusion of SDS/SES/racial adjustment in risk models; context-specific considerations for when they might be appropriate or inappropriate; strategies to avoid the potential adverse unintended consequences of such adjustment; concomitant monitoring for social and racial inequities through stratification; and special approaches for providers who care for high proportions of disadvantaged populations (e.g., payment adjustments, additional resources). Adjustment for SDS/SES/racial factors has generally been regarded as acceptable (e.g., in NQF white papers) when there is both an empirical association AND a plausible conceptual association of the risk variable with an outcome. For example, an SES/SDS/racial risk factor might be appropriate as a risk variable for readmission or mortality risk models, but not for CAUTI (catheter-associated urinary tract infections), CLABSI (central line-associated bloodstream infection), or process measures. For many outcomes, SES/SDS/racial adjustment is warranted to optimize risk model accuracy. For example, recent STS and Duke Clinical Research Institute analyses show that if race variables are excluded from some STS models, the resulting outcomes estimates are markedly different than the actual observed outcomes, and the O/E ratios are significantly different than unity, especially when the models are applied to racial minority subpopulations—in other

Topic	Commenter	Comment
		words, the models are less well calibrated, an essential feature of any risk model. This miscalibration persisted even when an SES/SDS indicator (specifically, dual eligible status) was simultaneously included in the models (i.e., thus addressing the hypothesis that the putative association of race and various outcomes is actually mediated by SES/SDS). Use of risk estimates from such models for patient counseling and shared decision-making would be misleading to patients and would inaccurately portray (and unfairly disadvantage) the risk-adjusted performance of providers, especially those caring for minority populations. Importantly, STS and its analytic center re-estimate risk factor coefficients several times annually, so that any changes in the association of race with outcomes will be implemented in the newest estimates. Further, STS is geocoding it adult cardiac surgery records and will use this information to derive an Area Deprivation Index for all patients with a valid address, thus providing us with the ability to further study the impact of race and SES/SDS using the what is arguably the most sensitive and comprehensive SES/SDS indicator. Finally, STS is aware of the recommendation in the ASPE report of October 2020 that functional status indicators be included in risk models as it may account for some of the impact on outcomes associated that is currently attributed to race. Although STS has a well-documented frailty indicator (5 meter walk test), it has not been collected with sufficient consistency by our participants to allow its inclusion in our models. Accordingly, STS has established a new working group on Frailty/functional indicators whose goal is to develop a new indicator that can be captured for virtually all patients using a combination of history, lab data, functional status, etc. Once developed, it will be added to STS models. Although SDS/SES/racial risk adjustment may be indicated to assure optimal risk model estimates based on current data, it is widely believed that such adjustment could potentially obscure disparities in care. To avoid this potential unintended consequence, most of the national guidance documents cited above recommend that any risk model results that are adjusted for SES/SDS/racial factors also present concomitant results in which outcomes are stratified by the same variables. This is a much more direct and explicit approach to monitor disparities and inequities and has been followed by STS in its risk modeling and performance measures. Please refer to the race-specific disparities data provided for each of the domains (mortality and morbidity) of measure 3030 under question 1b.4 (Importance tab) of the submission form (to be completed by the November submission deadline), which we believe will suffice to comply with this recommendation. STS Updates to Measure Testing Document Section 1.8 for Measures NQF #s 3030, 3031, 3032 - PART 2 1.8 What were the social risk factors that were available and analyzed? 1. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC, Jr., et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411-8.

Topic	Commenter	Comment
		2. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020. 3. National Quality Forum (NQF). Risk adjustment for Socioeconomic Status or other Sociodemographic Factors, accessed at http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx on June 24, 2020. 2014. 4. NQF. Evaluation of the NQF Trial Period for Risk Adjustment for Social Risk Factors. January 15, 2017. Available from: https://www.qualityforum.org/Publications/2017/07/Social_Risk_Trial_Final_Report.aspx. 5. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment. Washington, DC: The National Academies Press; 2017. 6. National Academies of Sciences, Engineering, and Medicine. Accounting for social risk factors in Medicare payment: Data. Washington, DC; 2016. 7. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Criteria, Factors, and Methods. Washington, DC: The National Academies Press; 2016. 8. National Academies of Sciences, Engineering, Medicine. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington, DC: The National Academies Press; 2016. 110 p. 9. Office of the Assistant Secretary for Planning and Evaluation USDoHaHS. Report to Congress: Social Risk Factors and Performance Under Medicare's Value-Based Purchasing Programs. A Report Required by the Improving Medicare Post-Acute Care Transformation (IMPACT) Act of 2014. Washington, DC; 2016. 10. 114th Congress of the United States. 21st Century Cures Act (Public Law 114–255). Washington, DC; 2016.