NATIONAL QUALITY FORUM

MEASURE WORKSHEET

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

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

Purple text represents the responses from measure developers.

Red text denotes developer information that has changed since the last measure evaluation review.

Brief Measure Information

NQF #: 3610

Corresponding Measures:

De.2. Measure Title: 30-day Risk Standardized Morbidity and Mortality Composite following Transcatheter Aortic Valve Replacement (TAVR)

Co.1.1. Measure Steward: America College of Cardiology

De.3. Brief Description of Measure: The TAVR 30-day morbidity/mortality composite is a hierarchical, multiple outcome risk model that estimates risk standardized results (reported as a "site difference") for the purpose of benchmarking site performance. This measure estimates hospital risk standardized site difference for 5 endpoints (death from all causes, stroke, major or life-threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation) within 30 days following transcatheter aortic valve replacement. The measure uses clinical data available in the STS/ACC TVT Registry for risk adjustment for the purposes of benchmarking site to site performance on a rolling 3-year timeframe.

1b.1. Developer Rationale:

S.4. Numerator Statement: A composite outcome including all-cause death, stroke, major or life-threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation within 30 days following transcatheter aortic valve replacement (TAVR).

If a patient experiences multiple outcomes captured in the overall rank composite measure, the outcome with the highest rank is assigned.

S.6. Denominator Statement: Patients who had TAVR.

S.8. Denominator Exclusions: Hospitals are excluded if they do not meet eligibility criteria noted in S.7.

Patients are excluded if any of the following occur:

1. They did not have a first-time TAVR in the episode of care (admission),
2. The TAVR was subsequent to another procedure in the Registry (other TAVR, Mitral Leaflet Clip and/or TMVR) during that admission.
3. The patient is readmitted for a repeat TAVR (re-admission) and the initial TAVR was performed during the rolling 3-year timeframe for the measure.

4. They are in TVT Registry sponsored research studies (identified with research study=yes and research study device used during procedure).

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Facility

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

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

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

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

Preliminary Analysis: New Measure

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention, or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

This composite measure estimates hospital risk-standardized site difference for five endpoints: (1) death from all causes, (2) stroke, (3) major or life-threatening bleeding, (4) acute kidney injury, and (5) moderate or severe paravalvular aortic regurgitation (PVL). The developers provided evidence for each outcome demonstrating actions a provider can take to achieve a change in the outcome.

Question for the Committee:

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

Guidance from the Evidence Algorithm

Assess performance on outcome (Box 1) à Relationship between outcome and healthcare action (Box 2) à Pass

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

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

Distribution of site-specific composite scores based on TAVR operations performed January 1, 2015 – December 31, 2017 (n = 52,561 records, 301 hospitals, data sources is the TAVR registry)

Mean	-0.004
Std Dev	0.037
Range	-0.16 to 0.06
IQR	-0.02 to 0.02
10%	-0.081
20%	-0.036
30%	-0.022
40%	-0.010
50%	-0.002
60%	0.005
70%	0.011
80%	0.018
90%	0.029
100%	0.049

Disparities

- Disparities data for individual endpoints is presented by race and ethnicity

Endpoints by race

Endpoint	White	Black	Asian	Other/Multiple
30-day Death	3.2% (1594/49458)	2.9% (53/1832)	2.5% (10/400)	2.6% (14/537)
30-day Stroke	2.5% (1225/49605)	2.6% (47/1842)	2.2% (9/401)	2.8% (15/541
Bleed	7.0% (3417/49054)	7.0% (127/1826)	7.0% (28/398)	8.8% (47/537)
Acute Kidney Injury	1.3% (621/49025)	2.1% (38/1840)	1.0% (4/395)	1.7% (9/539)
PVL	3.0% (1422/47695)	3.0% (53/1778)	3.9% (15/385)	2.7% (14/521)

Endpoints by ethnicity

Endpoint	Hispanic or Latino	Not Hispanic or Latino
30-day Death	3.2% (1607/50007)	3.0% (48/1601)
30-day Stroke	2.5% (1238/50160)	2.3% (37/1610)
Bleed	7.0% (3474/49608)	6.8% (109/1592)
Acute Kidney Injury	1.3% (646/49589)	1.3% (21/1595)
PVL	3.0% (1452/48278)	2.5% (38/1501)

Questions for the Committee:

Is there a gap in care that warrants a national performance measure?

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

1c. Composite – Quality Construct and Rationale

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

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

- This composite measure consists of two or more individual performance measure scores combined into one score.
- The endpoints were initially selected by an expert panel and then refined based on their adjusted association with 1-year mortality and the patient's quality of life as reported using the Kansas City Cardiomyopathy Questionnaire (KCCQ) summary score. The endpoints were then rank ordered based on the strength of the association. If a patient experiences more than one of the outcomes captured in the measure, the outcome with the highest rank is assigned. The developer uses a "win-ratio" analysis to arrive at an aggregation consistent with the association of the individual components with 1-year mortality and 1-year quality of life.

Questions for the Committee:

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

Is the method for aggregation and weighting of the components explicitly stated and logical?

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

Committee Pre-evaluation Comments:

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

1a. Evidence to Support Measure Focus: For all measures (structure, process, outcome, patient-reported structure/process), empirical data are required. How does the evidence relate to the specific structure, process, or outcome being measured? Does it apply directly or is it tangential? How does the structure, process, or outcome relate to desired outcomes? For maintenance measures –are you aware of any new studies/information that changes the evidence base for this measure that has not been cited in the submission? For measures derived from a patient report: Measures derived from a patient report must demonstrate that the target population values the measured outcome, process, or structure.

- Measure allows stratification of facilities based on five endpoints that have are likely of importance to patients considering the procedure.
- sufficient evidence for impactability
- Meets evidence requirements
- There is really good data to support this and a good logic model. Somewhat confused, and look forward to hearing form the measure developer, about the patient-reported outcomes included and alignment with description and numerator statement.
- support evidence is acceptable.
- This is a composite outcome measure. The relationship of the various components to the Kansas City Cardiomyopathy Questionnaire were tangential. CMS was requesting a composite that included self-report. It seems logical that the KCCQ score would be a component rather than related to a component
- Data apply directly to the outcomes and they are up to date
- Good Evidence
- Need more data.

1b. Performance Gap: Was current performance data on the measure provided? How does it demonstrate a gap in care (variability or overall less than optimal performance) to warrant a national performance measure? Disparities: Was data on the measure by population subgroups provided? How does it demonstrate disparities in the care?

- Some of the end points do vary by race and ethnicity, suggesting improvement capability or need to further assess patient differences.
- evidence for gap in care that warrants a national performance measure
- Yes, a performance gap is demonstrated
- Hard to determine the race/ethnicity differences with small denominators in non-white populations. Unknown quality with Range and IQR crossing zero.
- Good data but could be refined more if additional data was collected. overall for data collection, disparities needs to be reviewed in a broad way to determine how best to collect and use the data. This is not a measure by measure discussion at this point.
- The difference in composite scores seems very low (0.10 with a Std Dev of 0.037); I would say the performance gap is low. For disparities, several of the components were listed in relationship to race and ethnicity. The incidence of the components among the listed races and ethnicity were mostly less an 1% -- the highest difference in incidence was 1.8% for bleeds. Not sure this is clinically significant.
- There is a performance gap and moderate opportunity for improvement. Disparities are reported.
- Adequate evidence
- Need more data.

1c. Composite Performance Measure - Quality Construct (if applicable): Are the following stated and logical: overall quality construct, component performance measures, and their relationships; rationale and distinctive and additive value; and aggregation and weighting rules?

- End points are logical and explicitly stated.

- N/A
- Yes, for the most part
- rationale for construct seems appropriate.
- acceptable
- As mentioned before, I do not understand why the KCCQ was not one of the component measurements. The developer states that the components used have a demonstrated relationship to the KCCQ. I would rate the quality of the construct as low
- The quality construct is logical. I'm not clear on how the developers weight the 5 types of events
- Adequate
- Pennsylvania

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

2c. For composite measures: empirical analysis support composite approach

Reliability

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

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

Validity

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

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

Composite measures only:

2d. Empirical analysis to support composite construction. Empirical analysis should demonstrate that the component measures add value to the composite and that the aggregation and weighting rules are consistent with the quality construct.

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

Evaluators: NQF Scientific Methods Panel Subgroup

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel (SMP). The SMP Subgroup passed the measure on reliability and validity. The measure was not pulled for discussion during the March 2021 meeting. A summary of the measure and the SMP's review is provided below.

Reliability

Ratings for reliability: H-0; M-7; L-1; I-0; → Measure passes with MODERATE rating

Reliability testing was conducted at the measure score level:

- The developer estimated hospital-specific performance using a hierarchical proportional odds model on 100 sets of simulated data. Then, they calculated the Pearson correlation coefficient between each hospital's calculated estimate and the simulated true value. Reliability was calculated as the average squared Pearson correlation coefficient across the 100 data sets.
- The overall estimated reliability was 0.64, with a range from 0.65 for hospitals with at least 25 cases (n = 278) to 0.73 for hospitals with at least 200 cases (n = 96). The developer indicates they will be using a minimum of 60 cases over a three-year period for public reporting.
- In general, SMP subgroup members found the testing methodology appropriate and that the results supported moderate reliability.

Ratings for validity: H-3; M-5; L-0; I-0; → Measure passes with MODERATE rating

Validity testing was conducted at the composite measure score and component measure score level:

- The developer assessed the validity of the composite measure score using a known-group analysis. They divided the facilities into three levels of performance based on the global rank composite (i.e., better than expected, as expected, and worse than expected). Then, they examined the adjusted observed to expected (O/E) odds ratios for the individual components for each group. Sites with better-than-expected performance on the global rank composite metric showed lower O/E ratios when compared with sites that performed as expected or worse than expected. Sites that performed worse than expected showed consistently higher O/E ratios than other sites.
- The developer assessed the validity of the component measure scores using Cox proportional hazards modeling to evaluate the associations of the components with one-year mortality and average change in KCCQ-OS. All four non-fatal complications (components) were found to be associated with increased risk of one-year mortality and patient-reported health status (assessed via KCCQ-OS score).
- Exclusion of hospitals with more than 10 percent missing data for the global rank endpoint, baseline Kansas City Cardiomyopathy Questionnaire 12 (KCCQ-12) or baseline 5-meter walk test resulted in the exclusion of over half of the hospitals in the initial cohort (59,904 of 114,121).
- Covariates for case-mix adjustment were pre-selected based on inclusion in the risk model for NQF #3534 (TAVR 30-day mortality). Covariates were retained in the model regardless of their statistical significance. The developer did not collect or analyze any variables that directly measure social risk, based on the social risk analysis conducted for NQF #3534.
- C-statistic for Predicting an Outcome in One of the Worst Ranking Categories
  - Rank ≤ 1 = 0.70
  - Rank ≤ 2 = 0.65
  - Rank ≤ 3 = 0.63

- Rank ≤ 4 = 0.64
- Rank ≤ 5 = 0.63

The SMP subgroup members felt that the associations demonstrated through the analysis supported moderate to high validity.

Ratings for composite construction: H-3; M-3; L-1; I-1; → Measure passes with MODERATE rating

Composite construction:

- The global ranking endpoint is an ordinal categorical variable having six levels in which category one represents the worst possible outcome (death) and category six represents the best possible outcome (alive and free of major complications). Patients are classified according to the worst outcome (lowest rank score) that they experience. Endpoints were ranked in order of their decreasing hazard ratios with one-year mortality.
- The clinical importance of the complications was confirmed by assessing their associations with one-year mortality and one-year KCCQ-OS.
- The SMP sub-group members generally supported the composite construction. A couple of members questioned whether this measure represents a composite measure or a composite outcome and whether the additional complexity of this approach resulted in more precise measurement.

The SMP did not have any substantial concerns regarding the scientific acceptability of this measure.

Questions for the Committee regarding reliability:

Do you have any concerns that the measure can be consistently implemented (i.e., are measure specifications adequate)?

The Scientific Methods Panel is satisfied with the reliability testing for the measure. Does the Committee think there is a need to discuss and/or vote on reliability?

Questions for the Committee regarding validity:

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

The Scientific Methods Panel is satisfied with the validity analyses for the measure. Does the Committee think there is a need to discuss and/or vote on validity?

Questions for the Committee regarding composite construction:

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

The Scientific Methods Panel is satisfied with the composite construction. Does the Committee think there is a need to discuss and/or vote on the composite construction approach?

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

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

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

Committee Pre-evaluation Comments:

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

2a1. Reliability-Specifications: Which data elements, if any, are not clearly defined? Which codes with descriptors, if any, are not provided? Which steps, if any, in the logic or calculation algorithm or other specifications (e.g., risk/case-mix adjustment, survey/sampling instructions) are not clear? What concerns do you have about the likelihood that this measure can be consistently implemented?

- All are clearly defined.
- no concerns
- No concerns
- Aligned with SMP
- no concerns
- As a composite measure, this measure was evaluated by the Scientific Methods Panel. They deemed the reliability (Spearman 0.64) as moderate. I agree.
- This measure was reviewed by the Scientific Methods Panel (SMP). The SMP Subgroup passed the measure on reliability and validity.
- Defer to SMP, appears adequate
- Need more data.

2a2. Reliability - Testing: Do you have any concerns about the reliability of the measure?

- Measures seem reliable.
- no concerns
- As suggested, I agree that it would be preferable to have more recent data for testing
- Aligned with SMP
- no
- No
- The developer indicates they will be using a minimum of 60 cases over a three-year period for public reporting. The paradox of the measure is that the hospitals with the least experience (and perhaps poorest outcomes) are not reported.
- Adequate
- Need more data.

2b1. Validity - Testing: Do you have any concerns with the testing results?

- No concerns
- no concerns
- Though the overall reliability is OK, I share concerns about low volume hospitals
- Aligned with SMP
- no
- Reviewed by the Scientific Methods Panel. Validity rated as moderate. I agree
- I concur with the SMP
- Adequate
- Need more data.

2b4-7. Threats to Validity (Statistically Significant Differences, Multiple Data Sources, Missing Data) 2b4. Meaningful Differences: How do analyses indicate this measure identifies meaningful differences about quality? 2b5. Comparability of performance scores: If multiple sets of specifications: Do analyses indicate they produce comparable results? 2b6. Missing data/no response: Does missing data constitute a threat to the validity of this measure?

- No

- no concerns
- No substantial concerns
- Aligned with SMP
- all acceptable
- Scientific Methods Panel found no threats.
- Missing data are not a problem unless cases are not entered into the registry. All hospitals performing TAVR participate in the registry as a condition of CMS coverage.
- Using the mandated NCDR registry should make data quality good.
- Need more data.

2b2-3. Other Threats to Validity (Exclusions, Risk Adjustment) 2b2. Exclusions: Are the exclusions consistent with the evidence? Are any patients or patient groups inappropriately excluded from the measure? 2b3. Risk Adjustment: If outcome (intermediate, health, or PRO-based) or resource use performance measure: Is there a conceptual relationship between potential social risk factor variables and the measure focus? How well do social risk factor variables that were available and analyzed align with the conceptual description provided? Are all of the risk-adjustment variables present at the start of care (if not, do you agree with the rationale provided)? Was the risk adjustment (case-mix adjustment) appropriately developed and tested? Do analyses indicate acceptable results? Is an appropriate risk-adjustment strategy included in the measure?

- Agree
- adequate risk adjustment
- Agree with concerns from panel about lack of transparency in reporting SRF data
- Some concern with low numbers in non-White populations for risk adjustment sensitivity.
- all acceptable
- See Composite measure comments
- Exclusions and risk adjustment are appropriate.
- Exclusions for poor data quality may be expected to exclude poor quality programs, introducing bias. The exclusion rate was over 50% of records and over 40% of programs.
- Need more data.

2c. Composite Performance Measure - Composite Analysis (if applicable): Do analyses demonstrate the component measures fit the quality construct and add value? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

- Yes
- no concerns
- No concerns
- N/A
- Yes except that I feel that if they wanted a measure that included self-report, they should have directly incorporated the KCCQ score
- The event weights are derived from one-year mortality and one-year KCCQ-OS. This seems appropriate.
- Adequate
- Need more data.

Criterion 3. Feasibility

3. **Feasibility** is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
    - The data source for this measure is the STS/ACC TVT Registry. All hospitals performing TAVR participate in the registry as a condition of CMS coverage.

- The developer does not provide information on the fees required to use the registry.
- The necessary data are abstracted from a record by someone other than the person obtaining the original information (e.g., chart abstraction for registry).
- All of the data elements are in defined fields in electronic clinical data.

Questions for the Committee:

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

Are the required data elements available in electronic form, e.g., EHR or other electronic sources?

Is the data collection strategy ready to be put into operational use?

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

Committee Pre-evaluation Comments:
Criteria 3: Feasibility

3. **Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?**
   - No concerns
   - no concerns
   - No concerns
   - Good all data is in a registry and available to collect. Unsure on validity of data/data validation.
   - No concern
   - They are already being collected as part of the compulsory STS/ACC TVR Registry
   - The data source for this measure is the STS/ACC TVT Registry. All hospitals performing TAVR participate in the registry as a condition of CMS coverage.
   - Very Feasible
   - Need more data.

Criterion 4: Usability and Use

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

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

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

Current uses of the measure

Publicly reported? ☐ Yes ☒ No

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

OR

Planned use in an accountability program? ☒ Yes ☐ No

Accountability program details

The developer indicates that measure results will be voluntarily publicly reported on the STS Public Reporting Page by October 2021.

This measure is included in the Transcatheter Valve Certification for 2021, which fulfills the accountability program requirement.

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

Feedback on the measure by those being measured or others

1. Performance results are distributed to all TVT registry participants. These reports include detailed analysis of the institutions' performance in comparison to the registry population, an executive summary dashboard, and additional details to understand hospital performance within each composite category.
2. The developer noted that an initial open comment period was held during the development of this measure in 2019. Feedback was also obtained during a webinar to Registry Participants and other stakeholders during the question-and-answer session. Monthly registry site manager monthly calls, ad hoc phone calls and emails tracked with salesforce software, and registry-specific break-out sessions at the NCDR's annual meeting are held to provide on-going feedback.
3. The developer states that the feedback was generally positive and did not lead in a change in the measure.

Additional Feedback: none

***Questions for the Committee**:*

How can the performance results be used to further the goal of high-quality, efficient healthcare?

How has the measure been vetted in real-world settings by those being measured or others?

Preliminary rating for Use: ☒ Pass ☐ No Pass

4b. Usability (4a1. Improvement; 4a2. Benefits of measure)

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

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

Improvement results

The developer states the measure was first published to hospitals in the fall of 2020, therefore trends on performance improvement have not been analyzed.

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

Unexpected findings (positive or negative) during implementation

There were no unintended consequences to individuals or populations identified during testing or implementation.

Potential harms

None identified

Additional Feedback:

The developer states that sites have reported development of process improvement mechanisms and improvement in documentation practices as a result for TVT Registry implementation.

***Questions for the Committee**:*

Do the benefits of the measure outweigh any potential unintended consequences?

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

Committee Pre-evaluation Comments:
Criteria 4: Usability and Use

4a1. Use - Accountability and Transparency: How is the measure being publicly reported? Are the performance results disclosed and available outside of the organizations or practices whose performance is measured? For maintenance measures - which accountability applications is the measure being used for? For new measures - if not in use at the time of initial endorsement, is a credible plan for implementation provided? 4a2. Use - Feedback on the measure: Have those being measured been given performance results or data, as well as assistance with interpreting the measure results and data? Have those being measured or other users been given an opportunity to provide feedback on the measure performance or implementation? Has this feedback has been considered when changes are incorporated into the measure?

- Not stated
- no concerns, measure is part of accountability scheme
- Adequate
- Currently used in accountability programs and seems to be a part of public reporting
- a credible plan is provided. Great feedback.
- Feedback reported by the developer was positive
- I agree that feasibility is moderate
- Adequate
- Penn State/Hershey Medical Center

4b1. Usability – Improvement: How can the performance results be used to further the goal of high-quality, efficient healthcare? If not in use for performance improvement at the time of initial endorsement, is a credible rationale provided that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations? 4b2. Usability – Benefits vs. harms: Describe any actual unintended consequences and note how you think the benefits of the measure outweigh them.

- None stated
- no concerns
- Benefits likely outweigh harms
- Newer measure will little to none historical data to evaluate usability. Logical construct shows potential
- agree with document.
- Relatively new measure, so trends in performance improvement have not yet been analyzed
- It is currently used in an accountability program. It is not publicly reported. No harms have been identified.
- Composite model may not be as helpful in focusing improvement methods
- Need more data.

Criterion 5: Related and Competing Measures

Related or competing measures

NQF #3534 30-Day All-cause Risk Standardized Mortality Odds Ratio Following Transcatheter Aortic Valve Replacement (TAVR)
NQF #2561 STS Aortic Valve Replacement (AVR) Composite Score

Harmonization
The developer states that #3610 is fully harmonized with NQF #3534. NQF #3534 is one of the endpoints of #3610 and as such both measures address the same target population. Measure #3610's focus goes beyond that of NQF #3534 by including additional endpoints to mortality. Because both measures are calculated using the same registry data fields, there is no additional burden on providers for the collection or calculation of two separate measures. NQF #2561 has a similar focus but different target population than #3610. The developer outlines some differences in the specifications: population definition, events specific to each technique, timing of events, and risk model variables.

Committee Pre-evaluation Comments: Criterion 5: Related and Competing Measures

5. Related and Competing: Are there any related and competing measures? If so, are any specifications that are not harmonized? Are there any additional steps needed for the measures to be harmonized?

- Yes, no comment on harmonization
- no concerns, harmonized with related measures
- No concerns
- Will this replace some of the previous measures?
- No additional comments.
- #3534 - 30 day all-cause mortality from TAVR is a component of the current measure; 2561 STS is for the surgical population rather than transcatheter. The developer indicates harmonization to the extent possible
- NQF 3534 and 2561 are related but not competing
- No
- Need more data.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 06/10/2021

No NQF Members have submitted support/non-support choices as of this date.

No Public or NQF Member comments submitted as of this date.

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3610

Measure Title: 30-day Risk Standardized Morbidity and Mortality Composite following Transcatheter Aortic Valve Replacement (TAVR)

RELIABILITY: SPECIFICATIONS

Are submitted specifications precise, unambiguous, and complete so that they can be consistently implemented? ☒ Yes ☒ No

Submission document: "MIF_xxxx" document, items S.1-S.22

***NOTE**: NQF staff will conduct a separate, more technical, check of eCQM specifications, value sets, logic, and feasibility, so no need to consider these in your evaluation.*

Briefly summarize any concerns about the measure specifications.

Panel Member 1: None

Panel Member 4: No concerns

Panel Member 5: None

Panel Member 6: Evidence to support measure focus is blank on the MIF. Cannot assess that element. Submitter claims no data dictionary, but the STS registry has a very robust data dictionary. Submitting hospital may not be notified of adverse events that occur at another facility.

Panel Member 8: I was curious about the procedure date and home oxygen as risk adjusters. I can see why procedure year would be included in a 3-year model to capture secular tends, but the exact date seems like it has the potential to 'over fit' the model. On home oxygen, this is within provider control, so not an ideal risk adjuster. There is growing evidence that home oxygen is over prescribes. Since the registry has rich clinical detail, I'm not sure why this variable is necessary.

RELIABILITY: TESTING

Type of measure:

☒ Outcome (including PRO-PM) ☐ Intermediate Clinical Outcome ☐ Process

☐ Structure ☒ Composite ☐ Cost/Resource Use ☐ Efficiency

Data Source:

☐ Abstracted from Paper Records ☐ Claims ☒ Registry ☐ Abstracted from Electronic Health Record (EHR)
☐ eMeasure (HQMF) implemented in EHRs ☐ Instrument-Based Data ☐ Enrollment Data ☐ Other (please specify)

Level of Analysis:

☐ Individual Clinician ☐ Group/Practice ☒ Hospital/Facility/Agency ☐ Health Plan
☐ Population: Regional, State, Community, County or City ☐ Accountable Care Organization
☐ Integrated Delivery System ☐ Other (please specify)

Measure is:

☒ New ☐ Previously endorsed (NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.)

Submission document: "MIF_xxxx" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

Reliability testing level ☒ Measure score ☐ Data element ☐ Neither

Reliability testing was conducted with the data source and level of analysis indicated for this measure ☒ Yes ☒ No

If score-level and/or data element reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of **patient-level data** conducted?

☒ Yes ☒ No

Assess the method(s) used for reliability testing

Submission document: Testing attachment, section 2a2.2

Panel Member 1: Signal-to-noise analysis

Panel Member 2: Simulation. 100 sets of simulations of random sample of patients and estimation of the average correlation of score with true score.

Panel Member 4: This approach, which approximates a test-retest approach is appropriate for a risk-standardized measure where an SNR approach is less appropriate.

Panel Member 5: Correlation, as a reasonable approach

Panel Member 6: Data used for testing is from 1/1/2015 to 12/31/2017 - data is old in a clinical area that changes quickly. Interested in the clinicians' views on the age of the data.

Panel Member 8: The developers 'fit the hierarchical proportional odds model on 100 sets of simulated data having the same hospital-specific sample sizes and configurations of patient case mix as the actual analysis cohort'. They then looked at the correlation across these simulations. The method seems appropriate for the measure.

Panel Member 9: The developers fit a hierarchical proportional odds model on 100 sets of simulated data with the same hospital-specific sample sizes and patient case mix as the actual analysis cohort. In each of these simulated datasets, they implemented the measure methodology and calculated the Pearson correlation coefficient between each hospital's calculated estimate and the simulated true value. Because reliability may be impacted by the inclusion of hospitals with relatively few eligible patients, they also calculated correlation coefficients in subgroups of hospitals with at least 25, 50, 75, 100, or 200 eligible cases in the analysis.

Assess the results of reliability testing

Submission document: Testing attachment, section 2a2.3

Panel Member 1: The estimated reliability for the overall study cohort was 0.64, which indicates moderate reliability of the measure.

Panel Member 2: Average correlations were 0.64 for all hospitals, getting higher as sample restricted to hospitals with more cases, with correlation 0.73 for hospitals with more than 100 cases. Scatter plots and calibration analysis shows some underestimate of observed to expected for most serious outcomes when expected is high, and smaller overestimate of observed to expected when expected is low.

Panel Member 4: The results of the reliability testing are good, but use of 100 sets of simulated is a weakness as this may not be enough for the bootstrap approach, I'd expect to see at least 300 sets. Further, it would be helpful to see the confidence interval around the correlation.

Panel Member 5: Spearman-Brown, as a reasonable approach

Panel Member 6: Reliability testing breaks hospitals out by volume - it appears that hospitals with less than 25 cases may have a lower reliability - that value is not listed, but since all hospital reliability is 0.64 and reliability for hospitals with at least 25 cases is 0.65, I think we can infer that the reliability for the 23 hospitals with < 25 cases must be much lower that 0.65 to pull the overall number down.

Panel Member 8: The developers show an overall reliability score of 0.64, but find that as the hospital sample size goes up, so too does reliability. They have selected 60 cases as the minimum volume for a hospital to be included in the measure. Their table in 2a23 shows results for 50 and 75 cases, but not 60 -- would be helpful to see the results for 60 cases. Also, a 100-case minimum has a reliability score of 0.71 - that seems like a safer bet.

Panel Member 9: Reliability for the overall study cohort was 0.64 When the analytic cohort was restricted to sites with at least 25 cases over 3 years, reliability increased only slightly to 0.65. When the model is used in public reporting, the analysis will be restricted to sites that have at least 60 cases over a 3-year period. It would be good to know where the threshold of 60 came from because results are presented for 25, 50 (reliability 0.67), 75 (reliability 0,.69), 100 (reliability 0.71), and 200 cases (reliability 0.73)? Why 60? The developers state that the degree of correlation is "more than adequate" to justify the measures use in quality improvement and accountability. Leads to the question of whether the measure is sufficient if it were to be adopted for public reporting or value-based payment or other uses?

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

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☐ **No**

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

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

Submission document: Testing attachment, section 2a2.2

☒ **Yes**

☒ **No**

☒ **Not applicable** (data element testing was not performed)

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

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

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

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

☐ **Insufficient** (NOTE: Should rate **INSUFFICIENT** if you believe you do not have the information you need to make a rating decision)

Briefly explain rationale for the rating of OVERALL RATING OF RELIABILITY and any concerns you may have with the approach to demonstrating reliability.

Panel Member 1: The signal-to-noise analysis estimated reliability to be 0.64, which indicates moderate reliability.

Panel Member 2: Simulation approach shows reasonable if marginal level of correlation.

Panel Member 4: As noted above, I concern with the use of too few datasets for bootstrapping and the lack of the CI around the correlation coefficient.

Panel Member 5: Moderate levels demonstrated.

Panel Member 6: Reliability testing breaks hospitals out by volume - it appears that hospitals with less than 25 cases may have a lower reliability - that value is not listed, but since all hospital reliability is 0.64 and reliability for hospitals with at least 25 cases is 0.65, I think we can infer that the reliability for the 23 hospitals with < 25 cases must be much lower that 0.65 to pull the overall number down.

Panel Member 8: As mentioned above, by selection a 60-case minimum, they undercut the reliability a bit. As a result, I rated this measure moderate, not high.

Panel Member 9: Good reliability when measure has minimum number of cases, though over relatively long measurement period of 3 years which seems long?

VALIDITY: TESTING

Validity testing level: ☒ Measure score ☒ Data element ☐ Both

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

Submission document: *Testing attachment, section 2b1.*

☒ **Yes**

☒ **No**

☒ **Not applicable** (data element testing was not performed)

Method of establishing validity of the measure score:

☒ **Face validity**

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

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

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

Submission document: Testing attachment, section 2b1.

☒ **Yes**

☐ **No**

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

Assess the method(s) for establishing validity

Submission document: Testing attachment, section 2b2.2

Panel Member 1: Validity was assessed through the following associations: (1) Associations of component endpoints with 1-year mortality and KCCQ (2) Associations of component endpoints with categories based on overall composite (3) Assessment of case mix adjustment model

Panel Member 2: Correlation with one year mortality. Association with patient reported health status. Consistency of observed to expected for each outcome to overall ranking of better/as expected/worse.

Panel Member 4: The developer used evaluated the association between various endpoints in the composite with 1-year mortality and KCCQ change. The developer also looked at outcomes based on overall composite scores at the hospital level.

Panel Member 5: Components with outcomes and other components.

Panel Member 8: Among other things, the authors divided hospitals into groups by performance and compared the results on each of the major components in their composite score. Also, they show striking differences between 'better' and 'worse' performers, I this type of analysis more descriptive and validating.

Panel Member 9: Cox proportional hazards modeling was used to evaluate associations of short-term complications with one year mortality. Linear regression models were used to estimate the average change in KCCQ-OS associated with presence or absence of specific short-term complications. Covariates including age, gender, and multiple related comorbidities, along with complications were included simultaneously in both models to assess each complication, independent of other complications. To assess whether sites identified as

having high or low performance on the composite do not differ substantially with respect to one or more of the included endpoints, the developers used the concept of performance categories. Hospitals were labeled as having better than expected performance if the 95% probability interval surrounding their composite score (site difference) fell entirely above 0, as performing worse than expected if the 95% probability interval surrounding their composite score (site difference) fell entirely below 0, and as expected otherwise. They compared risk-adjusted mortality and complication rates for endpoints in the composite across the three defined performance groups. To assess the adequacy of the risk adjustment model to adjust for case mix, developers created calibration plots for the overall cohort and for several pre-specified subgroups. Large discrepancies between observed and predicted probabilities in any of the plots would suggest that the functional form of the model was misspecified and that estimates of provider performance may be invalid.

Assess the results(s) for establishing validity

Submission document: Testing attachment, section 2b2.3

Panel Member 1: Empiric validity analyses confirm the importance of the selected component endpoints by demonstrating their significant associations with one-year mortality and functional status. Agree with the developer that the test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance. These results support the validity of the composite measure as a quality measure for TAVR.

Panel Member 2: Validity seems to be justified. Metric based on proportions of compared comparisons may be difficult for providers to interpret/understand. Raw rates or adjusted of complications not clearly displayed, although appear in table in section 2b1.3.

Panel Member 4: The facility level results were the most compelling data to support the validity of the measure. Death, stroke, major bleeds, AKI and moderate/severe peri-valvular regurgitation all demonstrated the expected association with facility performance.

Panel Member 5: Significant associations. Empirical analyses confirm the importance of the selected endpoints by demonstrating their significant associations with one-year mortality and functional status. The test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance. These results support the validity of the composite measure as a quality measure for TAVR.

Panel Member 8: The developers show that the vast majority of cases land in the undifferentiated middle (the 'as expected cohort' with N=37,473). To the extent most hospitals are like each other on cardiac surgery, this suggests a degree of validation.

Panel Member 9: All 4 complications in the composite were found to be associated with increased risk of one year mortality, including patients with perioperative stroke (adjusted HR 2.10), major or life-threatening bleeding (adjusted HR 1.92), modified AKIN Stage III acute kidney injury (adjusted HR 1.81), and moderate or severe peri-valvular aortic regurgitation (adjusted HR 1.50). All were significant with $p<0.001$. Similarly, complications in the composite were also associated with 1-year patient reported health status as assessed by the KCCQ-OS score. Any stroke (adjusted impact -5.8 points) and moderate or severe paravalvular regurgitation (adjusted impact -2.0 points) were independently associated with poorer adjusted KCCQ-OS at one year. Other complications were not associated with one-year KCCQ-OS but were retained in the global rank composite measure, given their strong associations with 1-year mortality. Adjusted observed to expected (O/E) ratios of the individual endpoint components according to the 3 levels of site performance demonstrate that sites with better-than-expected performance on the global rank composite metric showed lower O/E ratios for all components of the global rank composite measure compared with the sites that performed as expected or worse than expected. Similarly, sites with worse than expected performance on the global rank composite demonstrated consistently higher O/E ratios than the other sites.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

Please describe any concerns you have with measure exclusions.

Submission document: Testing attachment, section 2b2.

Panel Member 1: None

Panel Member 2: Large number of sites excluded for incomplete data.

Panel Member 4: Exclusions seemed reasonable.

Panel Member 5: None

Panel Member 6: Concerned that low volume hospitals are not excluded.

Panel Member 8: No major concerns on exclusions.

Panel Member 9: Exclusions seem appropriate.

Risk Adjustment

Submission Document: Testing attachment, section 2b3

19a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

19b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

19c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☐ Not applicable

19c.2 Conceptual rationale for social risk factors included? ☒ Yes ☒ No

19c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☒ No

19d. **Risk adjustment summary:**

19d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

19d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion? ☒ Yes ☒ No

19d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☒ No

19d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration) ☒ Yes ☐ No

19d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☐ No

19e**. Assess the risk-adjustment approach**

Panel Member 1: Regarding inclusion of social risk factors (SRFs) including black race, other non-white race, Hispanic ethnicity, and participation in Medicaid, the developer mentioned that they tested the inclusion of SRFs in the testing phase but did not find statistical significance. Therefore, the developer decided not to capture that data or include in the model. However, it would have been more transparent to report this data in the testing document.

Panel Member 2: C-stats on risk adjustment model appear adequate. Not enough data presented on actual model results on social risk factors to assess their exclusion. Reference to statistical significance but not actual magnitude or impact on C-stats.

Panel Member 4: Social RFs beyond race were not included. Further, the risk adjustment model includes many theoretically relevant but statistically insignificant variables. C-statistics results show modest model performance.

Panel Member 5: Overall, fine. The consistent application of the model for similarity with another measures is one approach. The rationale in support of that approach could be stronger.

Panel Member 6: Hierarchical proportional odds model - appropriate in this case.

Panel Member 9: The global ranking endpoint is an ordinal categorical variable having 6 levels where category 1 represents the worst possible outcome (death) and category 6 represents the best possible outcome (alive and free of major complications). Variation in the distribution of outcome categories across hospitals was modeled by a hierarchical proportional odds model with hospital-specific random effects (intercepts). This multilevel hierarchical modeling framework was selected to estimate hospital-specific summary metrics while also adjusting for case mix and simultaneously accounting for the clustered data structure. Bayes methods were used to control for random fluctuations due to extreme outcomes. Covariates were pre-selected and were retained in the model regardless of their apparent statistical significance. The developers noted they did this because the statistical significance of a covariate's association with outcomes is partly a reflection of sample size and so a covariate with a non-significant p-value in the development sample could have a significant association in a future production run of the composite measure. Interesting to note that while they did not conceptual any association of social risk factors with the outcomes, the model covariates that were included could be directly or indirectly association with aspects of social risk, including age (age>75 OR=1.02, p-value)

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

Submission document: Testing attachment, section 2b4.

Panel Member 1: None

Panel Member 2: An average of 15% or so of patients encounter one or more of the complications in this composite. The better than expected have rates of 10% or so, the worse 25% or so, so substantial range in performance on this measure.

Panel Member 4: No concerns

Panel Member 5: None. Excluded sites were generally very similar to sites included in the analysis. Excluded sites had a numerically higher mean annualized volume but similar geographic distribution by region, urban versus rural setting and teaching versus non-teaching. Patients at included versus excluded sites were similar. The exclusion of a high proportion of sites with inadequate data completeness had minimal impact on the classification results of sites that met inclusion criteria for the measure.

Panel Member 6: NA

Panel Member 8: The table and histogram in section 2b4.2. is very helpful. It appears that this measure is good at identifying extreme outliners but does little for the vast majority of hospitals in the middle.

Panel Member 9: The identified differences in performance are both statistically significant and clinically meaningful.

Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified.

Submission document: Testing attachment, section 2b5.

Panel Member 1: None

Panel Member 2: NA

Panel Member 4: NA

Panel Member 5: NA

Panel Member 6: NA

Panel Member 8: Not applicable.

Please describe any concerns you have regarding missing data.

Submission document: Testing attachment, section 2b6.

Panel Member 1: The analysis was restricted to hospitals with ≥90% complete non-missing data for 30-day outcome status, baseline KCCQ-12 score, and baseline gait speed. I have no concern regarding these variables. However, it is not clear what was the proportion of missingness in covariates other than these three key variables, and how such missingness was handled.

Panel Member 2: While exclusion of sites with missing data does not seem to affect distribution for sites included, large number of exclusions makes measure less valuable.

Panel Member 4: Difficult to ascertain missing data in the sample due to the analyses presented (restricted to hospitals with hospitals with >=90% non-missing data.

Panel Member 5: None

Panel Member 6: Exclusions of for missing data removed ~50% of the patients and providers. Testing for impact of missing data included a number of demographics, but not the end point.

Panel Member 8: None.

Panel Member 9: Only 54% of hospitals with TAVR episodes were able to report performance scores which is problematic. (301/556)

For cost/resource use measures ONLY:

Are the specifications in alignment with the stated measure intent?

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

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

Panel Member 5: None

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

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

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

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

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

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

Panel Member 1: My overall rating is based on my validity sub-criteria assessments in #12 through #24.

Panel Member 2: Reasonable correlation of observed to expected and association with 1 year death and patient reported health.

Panel Member 3: See notes under composite

Panel Member 4: Outcomes of death and adverse events showed a strong relationship with measure score.

Panel Member 5: Approaches within and across composite to demonstrate associations was reasonable.

Panel Member 8: Evidence of validity provided by developers was helpful, but somewhat limited.

Panel Member 9: The empirical analyses confirm the selected endpoints have significant associations with one-year mortality and functional status.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

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

- ☒ **High**
- ☒ **Moderate**
- ☒ **Low**
- ☒ **Insufficient**

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

Panel Member 1: The component endpoints in the composite were selected on the basis of having a durable impact and being associated with outcomes that are most important to patients, such as longer-term mortality and quality of life.

Panel Member 2: Composite generally supported. Biggest variation is in Bleeding measure.

Panel Member 3: This is an excellent measure and should be rated high. But I disagree with treated it as a composite measure. Rather it is an outcome measure with a composite outcome. The validity evidence presented is best practice for establishing "data element" or "person or encounter" level validity. But the measure is not a composite measure, and the validity evidence is not related to composite measures.

Panel Member 4: The adjusted HR for 1 year mortality supports the endpoint ranking, however, it would be helpful to understand the confidence intervals of the HRs as they were fairly close.

Panel Member 5: Approaches within and across composite to demonstrate associations was reasonable.

Panel Member 6: Unclear that the added complexity of the composite measure results in a more precise comparison of performance.

Panel Member 8: This is where the developers did a particularly good job showing that all of the components moved in the same direction and had an empirical relationship to mortality.

Panel Member 9: The test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance. These results support the validity of the composite measure as a quality measure for TAVR.

ADDITIONAL RECOMMENDATIONS

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

Developer Submission

NQF #: 3610

Corresponding Measures:

De.2. Measure Title: 30-day Risk Standardized Morbidity and Mortality Composite following Transcatheter Aortic Valve Replacement (TAVR)

Co.1.1. Measure Steward: America College of Cardiology

De.3. Brief Description of Measure: The TAVR 30-day morbidity/mortality composite is a hierarchical, multiple outcome risk model that estimates risk standardized results (reported as a "site difference") for the purpose of benchmarking site performance. This measure estimates hospital risk standardized site difference for 5 endpoints (death from all causes, stroke, major or life-threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation) within 30 days following transcatheter aortic valve replacement. The measure uses clinical data available in the STS/ACC TVT Registry for risk adjustment for the purposes of benchmarking site to site performance on a rolling 3-year timeframe.

1b.1. Developer Rationale:

S.4. Numerator Statement: A composite outcome including all-cause death, stroke, major or life threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation within 30 days following transcatheter aortic valve replacement (TAVR).

If a patient experiences multiple outcomes captured in the overall rank composite measure, the outcome with the highest rank is assigned.

S.6. Denominator Statement: Patients who had TAVR.

S.8. Denominator Exclusions: Hospitals are excluded if they do not meet eligibility criteria noted in S.7.

Patients are excluded if any of the following occur:

1. They did not have a first-time TAVR in the episode of care (admission),
2. The TAVR was subsequent to another procedure in the Registry (other TAVR, Mitral Leaflet Clip and/or TMVR) during that admission.
3. The patient is readmitted for a repeat TAVR (re-admission) and the initial TAVR was performed during the rolling 3-year timeframe for the measure.
4. They are in TVT Registry sponsored research studies (identified with research study=yes and research study device used during procedure).

De.1. Measure Type: Composite

S.17. Data Source: Registry Data

S.20. Level of Analysis: Facility

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

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

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

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

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

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

1a. Evidence to Support the Measure Focus – See attached Evidence Submission Form

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

1a. Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 3610

Measure Title: 30-day Risk Standardized Morbidity and Mortality Composite following Transcatheter Aortic Valve Replacement (TAVR)

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

Date of Submission: 4/6/2021

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

Outcome

☐ Outcome:

- ☐ Patient-reported outcome (PRO):

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

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

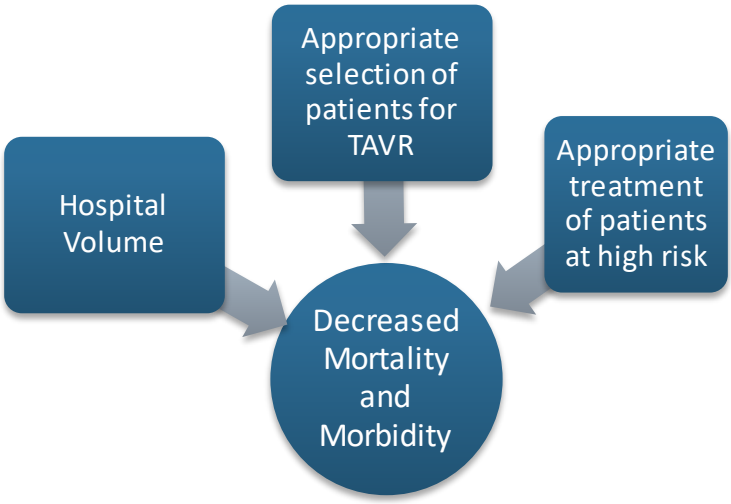
☐ Process:

- ☐ Appropriate use measure:

☐ Structure:

☒ Composite: 30-day Risk Standardized Morbidity and Mortality Composite following TAVR

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



We developed a risk standardized composite outcome measure for transcatheter aortic valve replacement (TAVR), which incorporated patient health status. The threefold goals of this outcome measure were to: benchmark performance for the purpose of quality-of-care monitoring; assist patients in their health care choices; and respond to CMS guidance [1]. The Centers for Medicaid and Medicare Services request the use of a composite model, which includes both fatal and non-fatal complications, as well as patient-reported health status to assess for quality (as compared to existing volume requirements).

The model's composite endpoints were ranked and chosen based on the association of these complications with late mortality as well as patient-reported health status. Any outcome with a significant hazard ratio was maintained in the model. Other endpoints (e.g., permanent pacemaker and vascular complications) were not included based on their lack of significance as it relates to 1-year mortality and 1-year Kansas City Cardiomyopathy Questionnaire (KCCQ) overall score, a well validated and proven tool for this patient population. [2,3]

References:

1. National Coverage Analysis (NCA) Tracking Sheet for Transcatheter Aortic Valve Replacement (TAVR) (CAG-00430R). Centers for Medicare and Medicaid Services Website. https://www.cms.gov/medicare-coverage-database/details/nca-tracking-sheet.aspx?NCAId=293 . Accessed on February 23, 2021.
2. Desai, N. D. (on behalf of the STS/ACC TVT Registry Risk Model Subcommittee). A Composite Metric for Benchmarking Site Performance in Transcatheter Aortic Valve Replacement: Results From the STS/ACC TVT Registry. Late-Breaking Clinical Trial at: ACC Scientific Session Together With World Congress of Cardiology (ACC.20/WCC). March 29, 2020.
3. Green CP, Porter CB, Bresnahan DR, Spertus JA. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. J Am Coll Cardiol. 2000 Apr;35(5):1245-55

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

Not applicable

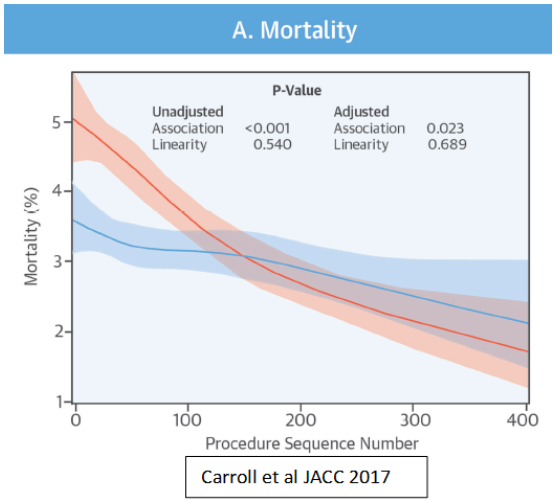
****RESPOND TO ONLY ONE SECTION BELOW -EITHER 1a.2, 1a.3 or 1a.4)****

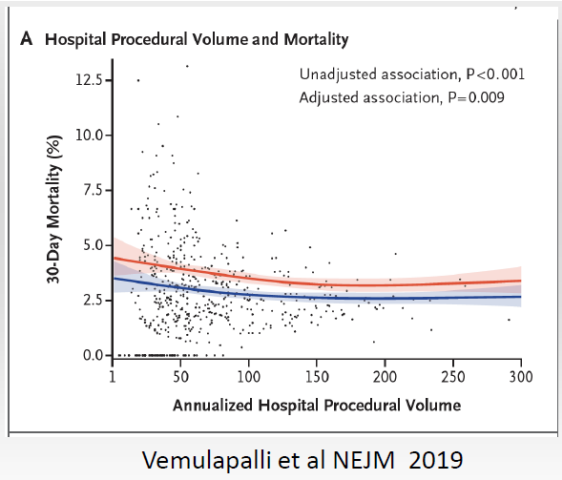
1a.2 FOR OUTCOME MEASURES including PATIENT REPORTED OUTCOMES - Provide empirical data demonstrating the relationship between the outcome (or PRO) to at least one healthcare structure, process, intervention, or service.

Mortality [1,6,8,9,10,11,12]

Mortality is arguably the worst outcome associated with an interventional procedure for patients. The published literature on TAVR demonstrates wide variation in TAVR mortality occurring at the hospital level [1,6,9,10]. Although greater procedural volume has been shown to correlate with better TAVR outcomes this correlation only accounts for a fraction of the variability in site performance [9,10]. Several studies demonstrate that there are other factors that can have a positive impact to reduce 30-day mortality. Besides hospital/operator volume (experience) these include appropriate patient selection and many aspects of post-procedure care.

The potential relationship between TAVR volume and outcomes, including mortality, was evaluated in two studies [9.10]. Carroll et al., [10] performed an analysis of TVT Registry data from 2011 to 2015 with 42,998 procedures performed at 395 hospitals. The authors found that lower volume was associated with higher in-hospital mortality as well as vascular complications, bleeding, and stroke, particularly with those hospitals with less than 100 procedures. Vemulapalli et al., [9] summarized the volume/outcome experience of the TVT Registry from 2015 to 2017 with 113,662 TAVR procedures performed at 555 hospitals by 2960 operators. The authors noted 30-day mortality was higher and more variable at hospitals with a low procedural volume as compared to hospitals with a high procedural volume. These results further validate the relationship between increased site volume with lower mortality rates (see figures below).





Stroke [13,14,15,16]

Stroke is the most debilitating complication after TAVR. Incidence of stroke has remained essentially the same over the past ten years and is associated with a high mortality rate. Physicians should be aware of patient risk factors that increase the risk of stroke (e.g., history of smoking, presence of peripheral arterial disease, avoidance of atrial fibrillation, and use of anticoagulants or antiplatelets) [13,14]. In addition, transcatheter cerebral embolic protection (TCEP) has been recently approved and may reduce the incidence of stroke. The literature has mixed results on the role of TCEP. However, clinical trials have demonstrated that the devices capture embolic debris in 99% of patients having TAVR and may reduce the incidence of stroke [15,16].

Acute Kidney Injury [17, 18, 19]

Assessing glomerular filtration rates pre- and post-procedure, identifying other pre-procedure risk factors for acute kidney injury (AKI), and adjusting the management of this high-risk group accordingly will reduce the incidence of AKI and mortality in patients who undergo TAVR [18]. Alsabbagh et al., [19] found hemodynamic monitoring approaches, composition of fluids in intravenous replacement therapy and avoidance of nephrotoxic agents, especially for patients deemed to be higher risk for AKI, can each result in lower incidence of AKI. Amin et al, [17] identified that contrast use varies among physicians and the amount of contrast used was not decreased for patients with higher risk of AKI. These findings identify opportunities to reduce AKI in patients who undergo TAVR.

Bleeding [20,21,22]

Bleeding complications are associated with worse short-term clinical outcomes including all-cause mortality [21]. In addition, Sherwood et al [22] notes that though patients should be discharged on dual antiplatelet therapy (DAPT) following TAVR, gaps in practice patterns persist and performance varied significantly among hospitals. Patients discharged with DAPT had a significantly higher risk for bleeding.

Identifying pre-operative risk factors for bleeding, understanding the benefits and risks of DAPT and adjusting the management of patients at risk for bleeding will lower the incidence of bleeding and improve 30-day mortality for patients who undergo TAVR. Hospitals should adopt bleeding avoidance strategies such as appropriate use of anticoagulants, determination of best access site and sheath size for the patient and use of closure devices. New developments such as reducing sheath sizes, alternative access sites and guidelines that optimize antithrombotic therapy are necessary to reduce the incidence of this detrimental complication.

Paravalvular Leak [23,24]

Mild residual paravalvular leak (PVL) is common and can be clinically silent after aortic valve replacement. Fortunately, more significant PVL is infrequent. Once detected, it must be monitored because it is associated with hemodynamic deterioration, and worse outcomes. It may require subsequent treatment.

Appropriate valve sizing pre-procedure is paramount to avoid paravalvular leakage after TAVR. In the past this was performed either before or during the procedure using an echo. More recently, pre-procedure CT imaging of the annulus area has become gold standard to perform more reliable assessment of AV annulus size, thus minimizing patient/prosthesis mismatch [24]

Additionally, procedure techniques can detect and treat PVL during and after TAVR procedures. Several scenarios have been identified as the most common causes of PVL and include [23]:

1. Incomplete adherence of the prosthesis to the aortic annulus due to severe native calcification causing incomplete or asymmetrical expansion of the frame. In these scenarios, post-dilation techniques appear very effective in reducing PVL.
2. Low implantation of the prosthesis or skirt below the annular ring. In these cases, the valve can be repositioned. In some cases, valve-in-valve techniques may be required.
3. High implantation of the prosthesis. This typically requires a 2nd prosthesis.

Heath Status [25,26]

Hospitals within the TVT Registry are mandated to assess patient health status, via the Kansas City Cardiomyopathy Questionnaire (KCCQ) for all patients before the procedure and in the 30 day and 1 year follow-up assessments. The KCCQ is a well validated and tested health status measurement tool, assessed directly from patients, that integrates 2 clinically relevant factors (symptoms and functional status) that may predict TAVR outcomes. Arnold (et al) examined whether a worse pre-procedure patient health status, as assessed by the KCCQ, was associated with greater long-term mortality after TAVR [25]. The authors concluded that patients with worse health status had a two-fold increased risk of death during the first year after TAVR, whereas those with poor and fair health status at baseline had intermediate outcomes. These results demonstrate that assessing patient functional status can support appropriate patient selection and accurately assess mortality risk for patients considering TAVR.

Gait Speed [27,28,29]

Gait speed is an independent predictor of both 30-day mortality and morbidity. This simple pre-procedure assessment helps identify frail patients who are at higher risk allowing hospitals and operators the ability to anticipate a higher level of post procedure care needs.

This TAVR measure is a composite of multiple outcomes that utilizes patient reported health status as a risk variable, for the purpose of hospital benchmarking as well as public reporting. The evidence supports multiple structures and processes exist to help sites improve their performance and improve outcomes of patients with TAVR.

General references

1. Desai, N. D. (on behalf of the STS/ACC TVT Registry Risk Model Subcommittee). A Composite Metric for Benchmarking Site Performance in Transcatheter Aortic Valve Replacement: Results From the STS/ACC TVT Registry. Late-Breaking Clinical Trial at: ACC Scientific Session Together With World Congress of Cardiology (ACC.20/WCC). March 29, 2020.
2. Carroll, JD., Mack, MJ, Vemulapalli, S., et al. STS-ACC TVT Registry State TAVR (State-of-the-art-review). JACC Vol. 76, No 21, 2020, pages 2492-2516.
3. Grover FL, Vemulapalli S, Carroll JD, et al. STS/ACC TVT Registry. 2016 Annual Report of The Society of Thoracic Surgeons/American College of Cardiology Transcatheter Valve Therapy Registry. J Am Coll Cardiol. 2017 Mar 14;69(10):1215-1230.

4. National Coverage Analysis (NCA) Tracking Sheet for Transcatheter Aortic Valve Replacement (TAVR) (CAG-00430R). Centers for Medicare and Medicaid Services Website. https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=293 Accessed on February 25, 2021.
5. Arnold SV, Manandhar P, Vemulapalli S, et al. Impact of Short-Term Complications of TAVR on Longer-Term Outcomes: Results from the STS/ACC Transcatheter Valve Therapy Registry. Eur Heart J Qual Care Clin Outcomes. 2020 Jan 11. pii: qcaa001. doi: 10.1093/ehjqcco/qcaa001
6. Murugiah, K, Wang, Y. Nihar R. et al. Hospital Variation in Outcomes for Transcatheter Aortic Valve Replacement Among Medicare Beneficiaries, 2011 to 2013. J Am Coll Cardiol. 2015 Dec, 66 (23) 2678–2679
7. Wayangankar, S. A., et al (2019). Length of Stay After Transfemoral Transcatheter Aortic Valve Replacement. JACC: Cardiovascular Interventions, 2019 Mar, 12(5), 422–430. doi: 10.1016/j.jcin.2018.11.015
8. Vemulapalli, S, et al. Hospital Resource Utilization Before and After Transcatheter Aortic Valve Replacement: The STS/ACC TVT Registry. JACC 73 (10) 2019, pg 1135-1146.

Procedure volume and outcomes

9. Vemulapalli S, Carroll JD, Mack MJ, et al. Procedural Volume and Outcomes for Transcatheter Aortic-Valve Replacement. N Engl J Med. 2019 Jun 27;380(26):2541-2550.
10. Carroll JD, Vemulapalli S, Dai D. Procedural experience for transcatheter aortic valve replacement and relation to outcomes. JACC. 2017;70:29–41.

Mortality

11. O'Brien SM, Cohen DJ, Rumsfeld JS, et al. Variation in Hospital Risk–adjusted Mortality Rates Following Transcatheter Aortic Valve Replacement in the United States: a Report from the Society of Thoracic Surgeons/American College of Cardiology Transcatheter Valve Therapy Registry. Circulation: Cardiovascular Quality and Outcomes. 2016 Sep;9(5):560-5.
12. Arnold SV, O'Brien SM, Vemulapalli S, et al. Inclusion of Functional Status Measures in the Risk Adjustment of 30-Day Mortality After Transcatheter Aortic Valve Replacement. JACC CI. 2018.

Stroke

13. Thourani, VH, O'Brien, SM, Kelly, JJ, et al. Development and Application of Risk Prediction Model for In-Hospital Stroke after TAVR: A Report From the STS/ACC TVT Registry. Ann Thorac Surg. 2019.
14. Huded CP, Tuzcu EM, Kishnaswamy A, et al. Association Between Transcatheter Aortic Valve Replacement and Early Postprocedural Stroke. JAMA Card. 2019.
15. Seeger, J., Gonska, B., Otto, M. Cerebral Embolic Protection During Transcatheter Aortic Valve Replacement Significantly Reduces Death and Stroke Compared With Unprotected Procedures. JACC Cardiovasc Interv. 2017 Nov 27;10(22):2297-2303.
16. Kapadia, S.R, Kodali, S., Makkar, R, et al. Protection Against Cerebral Embolism During Transcatheter Aortic Valve Replacement. JACC 2017 Jan 31; 69(4):367-377.

AKI

17. Amin, A.P., Bach, R. G, Caruso, M.L. Association of Variation in Contrast Volume With Acute Kidney Injury in Patients Undergoing Percutaneous Coronary Intervention JAMA Cardiol. 2017;2(9):1007-1012.
18. Hansen, JW, Foy, A, Yadav, P, et al. Death and Dialysis After Transcatheter Aortic Valve Replacement. An Analysis of the STS/ACC TVT Registry. JACC CI. Sept. 2017.
19. Alsabbagh MM, Asmar A, Ejaz NI, et al. Update on clinical trials for the prevention of acute kidney injury in patients undergoing cardiac surgery. Am J Surg 2013;206:86-95

Bleeding

20. Vora, A.N., Peterson, E.D., McCoy, L.A. The Impact of Bleeding Avoidance Strategies on Hospital-Level Variation in Bleeding Rates Following Percutaneous Coronary Intervention Insights From the National Cardiovascular Data Registry CathPCI Registry. JACC Interventions Vol 9, No. 8, April 25 2016, pages 771-9.
21. Sherwood MW, et al. Incidence, Temporal Trends, and Associated Outcomes of Vascular and Bleeding Complications in Patients Undergoing Transfemoral TAVR - Insights from the STS/ACC TVT Registry. Circulation: Cardiovascular Interventions. 15 Jan 2020.
22. Sherwood MW, Vemulapalli, S., Harrison, J.K.,et al. Variation in post-TAVR antiplatelet therapy utilization and associated outcomes: Insights from the STS/ACC TVT Registry. American Heart Journal. 2018 Oct: Vol 204, No 0, pages 9-16.

Paravalvular Leak:

23. Saia, F., Martinez, C., Gafoor, S. Long-Term Outcomes of Percutaneous Paravalvular Regurgitation Closure After Transcatheter Aortic Valve Replacement A Multicenter Experience. JACC Interventions Vol 8 No 5, 2015.
24. George, I., Gugliemetti, L.C., Bettinger, N. Aortic Valve Annular Sizing – Intraoperative Assessment versus Preoperative Multidetector Computed Tomography. Circulation: Cardiovascular Imaging. Volume 10, Issue 5, May 2017.

Patient Reported Health Status

25. Arnold, SV, Spertus, JA, Vemulapalli, S, et al. Association of Patient Reported Health Status With Long-Term Mortality After Transcatheter Aortic Valve Replacement. A Report from the STS/ACC TVT Registry. Circulation Cardiovascular Interventions. 2015; 8: e002875.
26. Green CP, Porter CB, Bresnahan DR, Spertus JA. Development and evaluation of the Kansas City Cardiomyopathy Questionnaire: a new health status measure for heart failure. J Am Coll Cardiol. 2000 Apr;35(5):1245-55

Gait Speed (5-meter walk)

27. Alfredsson J, Stebbins, A., Brennan, M. et al. Gait Speed Predicts 30-Day Mortality After Transcatheter Aortic Valve Replacement: Results From the STS/ACC TVT Registry. Circulation. 2016;133.
28. Afilalo, J., Kim, S., O'Brien, S. et al. Gait Speed and Operative Mortality in Older Adults Following Cardiac Surgery. JAMA Cardiology, 2016 Jun 1;1(3):314-21.
29. Afilalo, J., Eisenberg, M.J., Morin, J., et al. Gait Speed as an Incremental Predictor of Mortality and Major Morbidity in Elderly Patients Undergoing Cardiac Surgery. JACC Vol 56 (20), 2010. 1668-1677.

1a.3. SYSTEMATIC REVIEW(SR) OF THE EVIDENCE (for INTERMEDIATE OUTCOME, PROCESS, OR STRUCTURE PERFORMANCE MEASURES, INCLUDING THOSE THAT ARE INSTRUMENT-BASED) If the evidence is not based on a systematic review go to section 1a.4) If you wish to include more than one systematic review, add additional tables.

What is the source of the systematic review of the body of evidence that supports the performance measure? A systematic review is a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. It may include a quantitative synthesis (meta-analysis), depending on the available data. (IOM)

N/A for the composite measure

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

☐ Other systematic review and grading of the body of evidence (*e.g., Cochrane Collaboration, AHRQ Evidence Practice Center*)

☐ Other

Systematic Review	Evidence
Source of Systematic Review: Title Author Date Citation, including page number URL	*
Quote the guideline or recommendation verbatim about the process, structure or intermediate outcome being measured. If not a guideline, summarize the conclusions from the SR.	*
Grade assigned to the **evidence** associated with the recommendation with the definition of the grade	*
Provide all other grades and definitions from the evidence grading system	*
Grade assigned to the **recommendation** with definition of the grade	*
Provide all other grades and definitions from the recommendation grading system	*
Body of evidence: Quantity – how many studies? Quality – what type of studies?	*
Estimates of benefit and consistency across studies	*
What harms were identified?	*
Identify any new studies conducted since the SR. Do the new studies change the conclusions from the SR?	*

*cell intentionally left blank

_____

1a.4 OTHER SOURCE OF EVIDENCE

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

1a.4.1 Briefly SYNTHESIZE the evidence that supports the measure. A list of references without a summary is not acceptable.

1a.4.2 What process was used to identify the evidence?

1a.4.3. Provide the citation(s) for the evidence.

1b. Performance Gap

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

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

1b.1. Briefly explain the rationale for this measure *(e.g., how the measure will improve the quality of care, the benefits or improvements in quality envisioned by use of this measure)*

If a COMPOSITE *(e.g., combination of component measure scores, all-or-none, any-or-none), SKIP this question and answer the composite questions.*

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(**This is required for maintenance of endorsement**. Include mean, std dev, min, max, interquartile range, scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include.) This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

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

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. ***(This is required for maintenance of endorsement**. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included.) For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b1) under Usability and Use.*

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

1c. Composite Quality Construct and Rationale

1c.1. A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score.

For purposes of NQF measure submission, evaluation, and endorsement, the following will be considered composites:

- Measures with two or more individual performance measure scores combined into one score for an accountable entity.

- Measures with two or more individual component measures assessed separately for each patient and then aggregated into one score for an accountable entity:
- all-or-none measures (e.g., all essential care processes received, or outcomes experienced, by each patient);

1c.1. Please identify the composite measure construction: two or more individual performance measure scores combined into one score

1c.2. Describe the quality construct, including:

- the overall area of quality
- included component measures and
- the relationship of the component measures to the overall composite and to each other.

1c.3. Describe the rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually.

1c.4. Describe how the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale.

Candidate components under consideration for the TAVR 30-day composite model were initially chosen via expert consensus from a list of possible procedural and 30 – day complications of TAVR. Components of the composite were further refined by measuring the independent association of each putative complication with

1-year mortality and

1- year quality of life as measured by the Kansas City Cardiomyopathy Questionnaire summary score.

Since, by expert consensus, it was also thought to be important to include death within the proposed 30-day composite measure, we adopted a global rank composite framework to combine fatal and non-fatal endpoints. Since the individual components of the endpoint were felt to be of differing clinical significance, the relative "ranking" of each of the non-fatal endpoints was determined based on the strength of its individual association with 1-year mortality and 1-year quality of life. Death was then added to this ranking as the "highest" ranked endpoint. Thus, although each of the component endpoints were technically "weighted" equally, the ordinal ranking of the endpoints, in combination with a "win-ratio" analysis, allowed for aggregation consistent with the association of individual components with the patient-centric values of 1-year mortality and 1-year quality of life.

Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

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

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

S.1. Measure-specific Web Page *(Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials. Do not enter a URL linking to a home page or to general information.)*

STS/ACC does not have a measure specific webpage. However, the TVT Registry v2 data definitions are noted at the below link. https://www.ncdr.com/WebNCDR/docs/default-source/tvt-public-pagedocuments/coderdatadictionary_pdf-(1).pdf?sfvrsn=2

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

This is not an eMeasure **Attachment:**

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

No data dictionary **Attachment:**

S.2c. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

No, this is not an instrument-based measure **Attachment:**

S.2d. Is this an instrument-based measure (i.e., data collected via instruments, surveys, tools, questionnaires, scales, etc.)? Attach copy of instrument if available.

Not an instrument-based measure

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

S.3.2. For maintenance of endorsement, please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

S.4. Numerator Statement *(Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome) DO NOT include the rationale for the measure.*

***IF an OUTCOME MEASURE**, state the outcome being measured. Calculation of the risk-adjusted outcome should be described in the calculation algorithm (S.14).*

A composite outcome including all-cause death, stroke, major or life threatening bleeding, acute kidney injury, moderate or severe paravalvular aortic regurgitation within 30 days following transcatheter aortic valve replacement (TAVR).

If a patient experiences multiple outcomes captured in the overall rank composite measure, the outcome with the highest rank is assigned.

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

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

NUMERATOR:

The composite of outcomes is:

All-cause in-hospital or 30-day death:

1. Discharge status of deceased or
2. Follow-up status=deceased and date of difference between index procedure and death date is <=30 or
3. 30-day follow-up status=deceased, death date is missing, and difference between index procedure and follow-up assessment date is <=75 days.2

In-hospital or 30-day stroke:

1. In-hospital event=ischemic, hemorrhagic or undetermined stroke or
2. Follow-up event= ischemic, hemorrhagic or undetermined stroke and date of difference between index procedure and event date is <=30.

In-hospital or 30 Day VARC major or life-threatening disabling bleed:

1) In-hospital event=unplanned vascular surgery or intervention and decrease between pre procedure hemoglobin and the lowest post procedure hemoglobin is at least 3 g/dL or
2) In-hospital event=transapical related event, transaortic related event, bleeding at access site, hematoma at access site, retroperitoneal bleeding, gastrointestinal bleed, genitourinary bleed, other bleed, or hemorrhagic stroke and at least one of the following must be true:
    I. Decrease between pre procedure hemoglobin and the lowest post procedure hemoglobin is at least 3 g/dL or
    II. At least 2 units of RBC/whole blood transfused.
3) Discharge status of deceased with a vascular primary cause of death or
4) Follow-up event=major bleeding event or life-threatening bleeding and date of difference between index procedure and event date is <=30 or
5) Follow-up status of deceased and difference between index procedure and death date is <=30 days (or death date is missing, documentation includes a vascular primary cause of death, and difference between index procedure and follow-up assessment date is <=75 days).

In-hospital acute kidney injury stage III (AKI) or 30-day new requirement for dialysis:

1) In-hospital minimum increase of 300% between pre procedure hemoglobin and post procedure hemoglobin or
2) In-hospital minimum of 0.5 mg/dL absolute increase between pre procedure hemoglobin and post procedure hemoglobin and a minimum 4 mg/dL post procedure creatinine or
3) In-hospital or follow-up event = new requirement for dialysis and date of difference between index procedure and event date is <=30.

In-hospital or 30-day moderate or severe paravalvular leak:

1) In-hospital post procedure aortic paravalvular severity is moderate or severe (and no instance of follow-up aortic valve regurgitation of none or follow-up paravalvular regurgitation is none, mild, moderate, or severe and associated with latest follow-up echocardiogram date within 25-75 days of index procedure).
2) Follow-up aortic paravalvular severity is moderate or severe and associated with latest follow-up echocardiogram date within 25-75 days of index procedure.

1 Note: If a patient experiences multiple outcomes captured in the overall rank composite measure, the outcome with the highest rank is assigned.

2 Note on missing date of death: The <=75 day follow-up assessment timeframe was identified to be a clinically reasonable surrogate to capture a 30 day death if 30 day follow-up date of death was missing (this occurred in 0.9% of deceased records from January 2015 to December 2017). Sometimes a status of "deceased" is known and documented but the exact date of death is not available. In addition, we validated the accuracy of 30-day mortality in the TVT Registry by comparing Registry data linked CMS claims data from 2012-2015. Across 3.5 years, 99.6% of the 29,247 patient records had no discrepancy.

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

Patients who had TAVR.

S.7. Denominator Details *(All information required to identify and calculate the target population/denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets –*

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)

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

Population: Patients who had TAVR.

Timeframe: Rolling three years

Eligibility:

1) Eligibility at the hospital level:
    a. Acceptable "Data Quality Report (green or yellow)" data submissions for each quarter in the reporting period.
    b. >=90% completeness of the following items for all patient records in the rolling 3-year reporting period:
        i. Computed Baseline Kansas City Cardiomyopathy Questionnaire (a key risk model covariate) AND
        ii. Baseline 5-meter walk test (a key model covariate), AND
        iii. Event status/30 day follow-up (patients meet criteria for any endpoint or has some 30-day follow-up assessment at least 21 days after index procedure.
    c. At least 60 TAVR procedures
    d. Enrolled and submitted data prior to the rolling 3 year timeframe.
2) Eligibility at the patient level: Hospitalization for first-time TAVR procedure

S.8. Denominator Exclusions *(Brief narrative description of exclusions from the target population)*

Hospitals are excluded if they do not meet eligibility criteria noted in S.7.

Patients are excluded if any of the following occur:

1) They did not have a first-time TAVR in the episode of care (admission),
2) The TAVR was subsequent to another procedure in the Registry (other TAVR, Mitral Leaflet Clip and/or TMVR) during that admission.
3) The patient is readmitted for a repeat TAVR (re-admission) and the initial TAVR was performed during the rolling 3-year timeframe for the measure.
4) They are in TVT Registry sponsored research studies (identified with research study=yes and research study device used during procedure).

S.9. Denominator Exclusion Details *(All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at S.2b.)*

1) Hospital ineligibility:
    a. Unacceptable data quality report submissions for all quarters of the reporting time-period.
    b. Hospitals who have less than 90% of patient records with respect to ANY of the following assessments in the rolling 3-year reporting period:
        i. Computed Baseline Kansas City Cardiomyopathy Questionnaire (a key risk model covariate) AND
        ii. Baseline 5-meter walk test (a key model covariate), AND

iii. Event status/30-day follow-up (patient meets criteria for any endpoint or 30-day follow-up assessment is performed at least 21 days after index procedure).

c. At least 60 TAVR procedures.

d. Enrolled and submitted data prior to the rolling 3 year timeframe.

2) Patient ineligibility:

a. They did not have a first-time TAVR in the episode of care (admission),

b. The TAVR was subsequent to another procedure in the Registry (other TAVR, Mitral Leaflet Clip and/or TMVR) during that admission.

c. The patient is readmitted for a repeat TAVR (re-admission) and the initial TAVR was performed during the rolling 3-year timeframe for the measure.

d. The patient is in a TVT Registry sponsored research studies (identified with research study=yes and research study device used during procedure).

S.10. Stratification Information *(Provide all information required to stratify the measure results, if necessary, including the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format with at S.2b.)*

In theory, estimates of provider-specific performance within specific disadvantaged patient populations (e.g., by race, ethnicity) could be generated by applying the measure´s modeling methodology to an analysis cohort that is restricted to members of the population of interest. As a practical matter, the number of patients per provider that belong to such populations may be too small to permit a meaningful comparison of performance across providers for these groups. Outcome disparities by race and ethnicity could potentially be assessed by including race and ethnicity in the risk adjustment model and reporting their odds ratios.

S.11. Risk Adjustment Type (Select type. Provide specifications for risk stratification in measure testing attachment)

Statistical risk model

If other:

S.12. Type of score:

Other (specify):

If other: Site difference

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

Better quality = Higher score

S.14. Calculation Algorithm/Measure Logic (*Diagram or describe the calculation of the measure score as an ordered sequence of steps including identifying the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period for data, aggregating data; risk adjustment; etc.)*

The measure score is calculated based on the following steps:

A. Patient cohort is identified based on inclusion criteria for a rolling-3-year time period (see questions S.7-S11)

B. Data elements for risk adjustment variables are analyzed using the first collected value (model variables listed below)

C. Observed and expected outcomes are ascertained for each hospital.

D. A measure score is calculated with aggregated data across all included sites. Case mix adjustment is implemented using a hierarchical logistic regression model with the above covariates and a site-specific random intercept.

   a. The main summary measure of a hospital's risk-standardized outcomes performance is the hospital's estimated "site difference" which calculates the probability that a random patient at the hospital of interest would have a worse outcome at an average hospital (vs the hospital of interest) MINUS the probability that a random patient at the hospital of interest would have a better outcome at an average hospital (vs the hospital of interest).

      i. A site difference >0 (positive number) implies that a random patient is better off at your hospital (vs an average hospital). This implies that a hospital has better than expected performance.

      ii. A site difference <0 (negative number) implies that a random patient is better off at an average hospital (not your hospital). This implies that a hospital has worse than expected performance.

   b. A 95% empirical Bayes interval is estimated for each facilities performance.

Risk adjustment variables include:

1. Age
2. Body surface area (BSA)
3. Sex
4. Race/ethnicity
5. Estimated glomerular filtration rate (eGFR), which quantifies kidney function
6. Left ventricular ejection fraction (LVEF)
7. Hemoglobin function
8. Platelet count
9. Procedure date
10. Dialysis
11. Left main coronary artery stenosis >=50%
12. Proximal left anterior descending coronary artery stenosis >=70%
13. Priori myocardial infarction
14. Endocarditis
15. Prior stroke or transient ischemic attack
16. Carotid stenosis
17. Prior peripheral artery disease
18. Current/recent smoker
19. Diabetes
20. Hypertension
21. Atrial fibrillation/flutter
22. Conduction defect
23. Severe chronic lung disease
24. Home oxygen
25. "Hostile" chest

26. Porcelain (severely concentrically calcified) aorta
27. Access site
28. Pacemaker
29. Previous implantable cardioverter defibrillator
30. Prior percutaneous coronary intervention
31. Prior coronary artery bypass surgery
32. # prior cardiac operations
33. Prior aortic valve surgery/procedure
34. Prior other valve surgery/procedure (mitral, tricuspid, pulmonic)
35. Aortic valve disease etiology
36. Aortic valve morphology
37. Aortic insufficiency (moderate or severe)
38. Mitral insufficiency (moderate or severe)
39. Tricuspid insufficiency (moderate or severe)
40. Acuity status (defined by a combination of procedure status, prior cardiac arrest w/in 24 hours, need for pre-procedure inotropic medications, and use of mechanical assist device)
41. Unable to walk
42. Gait speed (via the 5-meter walk test which assesses frailty)
43. Baseline Kansas City Cardiomyopathy Questionnaire-12 (KCCQ-12, a measure of heart-failure specific health status)

What is a Site Difference? A site difference assesses the association between risk factors and composite outcomes. It calculates the probability that a random patient at the hospital of interest would have a worse outcome at an average hospital (vs the hospital of interest) MINUS the probability that a random patient at the hospital of interest would have a better outcome at an average hospital (vs the hospital of interest).

It conceptually allows for a comparison of a particular hospital's performance given its case mix to an average hospital's performance with the same case mix. Thus, a lower site difference (<0) implies worse-than-expected morbidity/mortality (worse quality), and a higher site difference (>0) implies better-than-expected morbidity/mortality (better quality). To assess hospital performance in any reporting period, we re-estimate the model coefficients using the years of data in that period.

References:

a. Win Ratio–An Intuitive and Easy-To-Interpret Composite Outcome in Medical Studies: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5518256/
b. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials: https://pubmed.ncbi.nlm.nih.gov/10399200/
c. Use of the Win Ratio in Cardiovascular Trials – JACC Heart Failure https://www.jacc.org/doi/full/10.1016/j.jchf.2020.02.010

Normand S-LT, Shahian DM, 2007. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Stat Sci 22 (2): 206-226.

S.15. Sampling *(If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.)*

IF an instrument-based performance measure (e.g., PRO-PM), identify whether (and how) proxy responses are allowed.

N/A. This measure is not based on a sample or survey. Data from all hospitals and all TAVR procedures would be included in the process of re-estimating model variables.

S.16. Survey/Patient-reported data *(If measure is based on a survey or instrument, provide instructions for data collection and guidance on minimum response rate.)*

Specify calculation of response rates to be reported with performance measure results.

N/A. This measure is not based on a survey or patient-reported data.

S.17. Data Source *(Check ONLY the sources for which the measure is SPECIFIED AND TESTED).*

If other, please describe in S.18.

Registry Data

S.18. Data Source or Collection Instrument *(Identify the specific data source/data collection instrument (e.g., name of database, clinical registry, collection instrument, etc., and describe how data are collected.)*

IF instrument-based, identify the specific instrument(s) and standard methods, modes, and languages of administration.

STS/ACC TVT Registry

S.19. Data Source or Collection Instrument *(available at measure-specific Web page URL identified in S.1 OR in attached appendix at A.1)*

Available at measure-specific web page URL identified in S.1

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

Facility

S.21. Care Setting *(Check ONLY the settings for which the measure is SPECIFIED AND TESTED)*

Inpatient/Hospital

If other:

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

This composite includes both fatal and non-fatal outcomes that were ranked (death, stroke, major/life threatening bleed, acute kidney injury, moderate or severe paravalvular aortic regurgitation) and assigned a different weight based on their severity. These outcomes were selected, and rank ordered based on their independent association with 1-year mortality and the patient's quality of life (via KCCQ). Outcomes with a significant hazard ratio was maintained and outcomes that were not found to be significant (e.g., permanent pacemaker and vascular complication) were not included in the model.

2. Validity – See attached Measure Testing Submission Form

TAVR_Composite_model_testing_12_22_20.docx

2.1 For maintenance of endorsement

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

2.2 For maintenance of endorsement

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

2.3 For maintenance of endorsement

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

Measure Testing (subcriteria 2a2, 2b1-2b6)

Measure Number (*if previously endorsed*)**:** TBD
Composite Measure Title: 30-day Risk Standardized Morbidity and Mortality Composite following Transcatheter Aortic Valve Replacement (TAVR)
Date of Submission: **12/23/2020**

Composite Construction:

☒Two or more individual performance measure scores combined into one score

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

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

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

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

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☐ claims	☐ claims
☒ registry	☒ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other:	☐ other:

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

1.3. What are the dates of the data used in testing? Jan 1, 2015 – Dec 31, 2017

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

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

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

The final cohort included 52,561 records from 301 hospitals. Measure development and testing was based on TVT data from patients undergoing TAVR as the first recorded cath lab visit of their hospitalization who were discharged during Jan 1, 2015 – Dec 31, 2017. Only the first TAVR per patient during the 3-year period was included. Based on conventions established for the TVT 30-day mortality model, data from hospitals with >10% missing data for the outcome variable and other key study variables were excluded. From a starting population of 114,121 TAVR records from 556 TVT hospitals, we excluded hospitals with >10% missing data for the global rank endpoint, baseline KCCQ-12, or baseline 5-meter walk leaving 54,217 records from 301 hospitals. Finally, we excluded 1,656 records (3%) with missing data for the global rank endpoint.

Table: Characteristics of TVT TAVR sites in the development sample (2015-2017)

Annual TAVR volume, median (IQR)	58.7 (41.7, 90.9)
Teaching hospital, n (%)	169 (56.1)
Region, n (%)	*
South	113 (37.5)
Midwest	82 (27.2)
West	64 (21.3)
Northeast	41 (13.6)
Missing	1 (0.3)
Location, n (%)	*
Rural	33 (11.0)
Suburban	77 (25.6)
Urban	191 (63.5)
No. of Patient Beds, median (IQR)	438 (317, 608)

*cell intentionally left blank

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

Details of the cohort inclusion/exclusion identification are provided in Section 1.5 above. The final development cohort was 52,561 records from 301 hospitals. Patient pre-procedural characteristics are summarized in the following table.

Characteristic	Results*
Age, years	82 (76, 87)
Male sex, %	53.5 %
Diabetes mellitus, %	38.4 %
Current smoker, %	5.8 %
Currently on dialysis, %	3.5 %
GFR, ml/min/1.73 m2	63 (48, 78)
LVEF, %	58 (50, 60)
Prior MI, %	22.8 %
Prior pacemaker, %	14.2 %
Prior PCI, %	34.3 %
Prior CABG, %	23.4 %
Prior aortic valve procedure, %	11.9 %
Prior non-aortic valve procedure, %	2.2 %
NYHA Class IV, %	14.8 %
Atrial fibrillation/flutter, %	39.2 %
Conduction defect, %	38.0 %
Prior stroke or TIA, %	18.5 %
Carotid stenosis, %	21.8 %
Peripheral arterial disease, %	28.2 %
Severe chronic lung disease, %	11.8 %
Home oxygen, %	10.2 %
Hostile chest, %	7.0 %
Porcelain aorta, %	4.0 %
Non-femoral access, %	7.6 %
Acuity, elective	90.8 %
KCCQ-OS	42.7 (26.0, 62.5)
5MWT, seconds	7.3 (6.0, 9.7)

* Continuous variables reported as medians with 25th and 75th percentiles

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

Statistical model development and testing

Development and testing of the statistical model was based on TVT TAVR data from 2015-2017 as described in Section 1.5 above.

Selection and ranking of individual endpoints

Analyses to inform and validate the selection and ranking of component endpoints in the composite were based on a separate smaller TVT TAVR cohort that had been linked with Medicare data in order to follow patients longitudinally. The cohort included patients aged 65 or older who underwent transfemoral TAVR between January 1, 2015 and January 31, 2016 and were alive 30 days post TAVR, had records linked to Medicare, and had non-missing 30 day data (N=12,607). All of these records had at least 1 year of follow-up data. To evaluate associations between non-fatal periprocedural complications and 1 year KCCQ-OS, we included records with non-missing KCCQ-OS (N=10,883) and used inverse probability weighting to adjust for differences between patients who had missing versus non-missing KCCQ-OS.

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

Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic (SES) or sociodemographic (SDS) factors (e.g., race, ethnicity, education, income, payer [e.g., Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate [1]. 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. They also note that readily available SES factors have often not demonstrated significant impact on outcomes. As part of an NQF pilot project, STS specifically studied dual eligible status in the STS readmission measure [2] and found minimal impact. Finally, even proponents of inclusion of SES in risk models agree that these factors make more sense intuitively for some outcomes (e.g., readmission) than others (hospital mortality, complications)—that is, they are context-specific [2,3].

In identifying a risk adjustment approach for this measure, and in keeping with the general approach taken for the current risk models by the Society for Thoracic Surgeons [3], we chose to avoid the more philosophical and downstream health policy implications of SES adjustment and based our modeling decisions on empirical findings and consideration of the model's primary intended purpose--to adjust for case mix. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across TVT participants. For example, race and ethnicity will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator. Race has an empirical association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (e.g., genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension).

For the purposes of this NQF submission and the question on whether social risk factors should be included in risk adjustment, we modeled this composite to match the covariates used in NQF #3534, 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR). During testing of that measure, we examined variables for black race, other non-white race, Hispanic ethnicity, and participation in Medicaid and whether any of these variables had any statistically significant associations with

30-day mortality after adjusting for other factors in the hierarchical model. For each variable in each time period, the 95% confidence interval around the odds ratio overlapped with the null value of 1.0, which implies that there was no statistically significant association. As a result, we did not collect or analyze any variables that directly measure social risk in this composite beyond what is already included. The model covariates that are presumed to be directly or indirectly associated with aspects of social risk include age, sex, race, and ethnicity, among others.

1. National Academies of Sciences E, and Medicine. Accounting for social risk factors in medicare payment. Washington, DC: The National Academies Press; 2017.
2. Shahian DM, He X, O'Brien SM et al. Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. Circulation 2014;130(5):399-409.
3. Shahian DM, Jacobs JP, Badhwar V, 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 May;105(5):1411-1418.

2a2. RELIABILITY TESTING

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

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

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

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

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

Mathematically, reliability is the square of the correlation coefficient between a measurement and the true value. The main drivers of reliability are the hospital-specific sample sizes and the magnitude of true difference across the hospitals. In order to estimate reliability, we asked the following question. Suppose that hospital outcomes were accurately described by a proportional odds model identical to the one used for analysis and suppose that the true fixed effect parameter values generating the data were exactly equal to their estimates from the current analysis. If we re-estimated each hospital's performance in a different random sample of patients having the same hospital-specific case mix and sample sizes, how closely would the hospital-specific performance estimates agree with the underlying true values. To answer this question, we fit the hierarchical proportional odds model on 100 sets of simulated data having the same hospital-specific sample sizes and configurations of patient case mix as the actual analysis cohort. In each of these simulated datasets, we implemented the measure methodology and calculated the Pearson correlation coefficient between each hospital's calculated estimate and the simulated true value. Reliability was calculated as the average squared Pearson correlation coefficient across the 100 synthetic data sets. Because reliability may be impacted by the inclusion of hospitals with relatively few eligible patients, we also calculated correlation coefficients in subgroups of hospitals having at least 25, 50, 75, 100, or 200 eligible cases in the analysis.

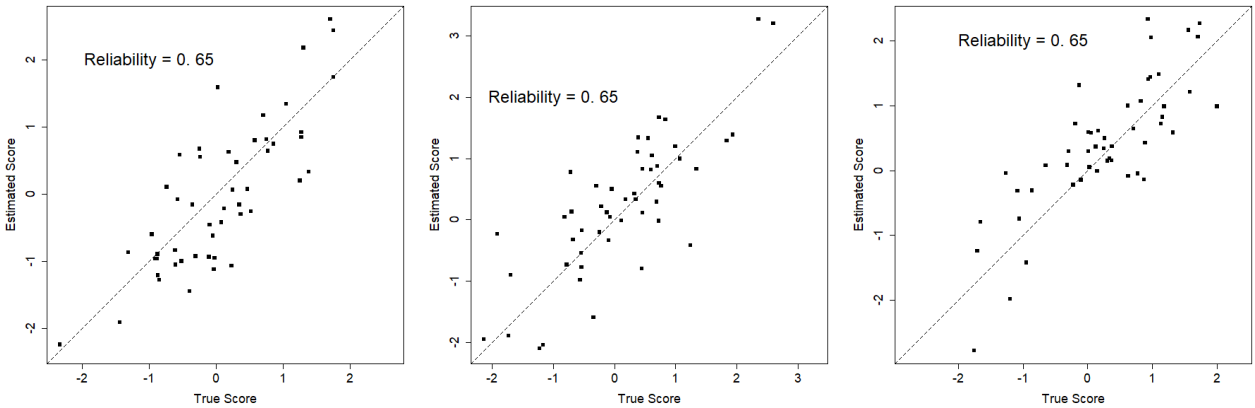
2a2.3. What were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

As shown in the table, estimated reliability for the overall study cohort was 0.64. When the analytic cohort was restricted to sites with at least 25 cases over 3 years, reliability increased to 0.65. When the model is used in public reporting, the analysis will be restricted to sites that have at least 60 cases over a 3-year period.

Hospital Data	Number of Hospitals	Estimated Reliability
All hospitals	301	0.64
Hospitals with at least 25 cases	278	0.65
Hospitals with at least 50 cases	248	0.67
Hospitals with at least 75 cases	219	0.69
Hospitals with at least 100 cases	187	0.71
Hospitals with at least 200 cases	96	0.73

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

The following 3 figures illustrate hypothetical scenarios in which the squared correlations between true and estimated values for a performance measure are exactly equal to 0.65, i.e., equal to the value we estimated for the TAVR composite reliability among hospitals with at least 25 cases. It's clear that the relationship between true and estimated values isn't perfect. However, it's also clear that sites with very low estimated scores are very likely to be on the low side of true performance and sites with very high estimated scores are very likely to be on the high side of true performance. We think this degree of correlation is more than adequate to justify the measure's use in quality improvement and accountability applications.



2b1. VALIDITY TESTING

Note*: Current guidance for composite measure evaluation states that validity should be demonstrated for the composite performance measure score. If not feasible for initial endorsement, acceptable alternatives include assessment of content or face validity of the composite OR demonstration of validity for each component. Empirical validity testing of the composite measure score is expected by the time of endorsement maintenance.*

2b1.1. What level of validity testing was conducted?

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

☒ **Composite performance measure score**

☒ **Empirical validity testing**

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

☒ **Validity testing for component measures** *(check all that apply)*

***Note**: applies to ALL component measures, unless already endorsed or are being submitted for individual endorsement.*

☐ **Endorsed (or submitted) as individual performance measures**

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

☒ **Empirical validity testing of the component measure score(s)**

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

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

Associations of component endpoints with 1-year mortality and KCCQ

Endpoints in the composite were selected on the basis of having a durable impact and being associated with outcomes that are most important to patients, such as longer-term mortality and quality of life. To confirm the importance of the selected short-term complications, we used data from a separate smaller TVT TAVR cohort that had been linked with Medicare data in order to follow patients longitudinally (see Section 1.7 for details). Cox proportional hazards modeling was used to evaluate associations of short-term complications with one year mortality. Linear regression models were used to estimate the average change in KCCQ-OS associated with presence or absence of specific short term complications. All complications were included simultaneously in both the mortality and KCCQ-OS models to assess each complication, independent of other complications. In addition to complications, covariates in the model included: Age, Sex, Body Surface Area, LVEF, Hemoglobin, Platelet count, GFR, Dialysis, Race White, Hispanic or Latino Ethnicity, Left Main Stenosis>=50%, Proximal LAD >=70%, Prior MI, Endocarditis, Prior Stroke/TIA, Carotid Stenosis, Prior PAD, Smoker, Diabetes, NYHA, Atrial Fibrillation, Conduction Defect, Chronic Lung Disease, Home Oxygen, Hostile Chest, Porcelain Aorta, Non-Femoral Access Site, Pacemaker, Previous ICD, Prior PCI, Prior CABG, Previous cardiac surgeries, Prior Aortic Valve procedure, Prior Non-Aortic Valve procedure, Degenerative AV Etiology, Tricuspid Valve Morphology, Aortic Valve Insufficiency, Mitral Valve Insufficiency, Tricuspid Valve Insufficiency and Acuity.

Associations of component endpoints with categories based on overall composite

When endpoints in a composite occur with different frequencies, the composite may tend to be heavily influenced by the relatively more frequent endpoints, which can be problematic if the relatively more frequent endpoints have lesser clinical importance and if the less important endpoints dominate. In the case of the TVT TAVR composite, there is not a major concern about an endpoint of lesser clinical importance dominating because all included endpoints have established associations with 1-year outcomes and the frequencies of the different endpoints are mostly of the same order of magnitude. However, it may still be concerning if sites identified as having high or low performance on the composite do not differ substantially with respect to one or more of the included endpoints. To assess this, we used the concept of performance categories to be more formally introduced in 2b4. Hospitals were labeled as having better than expected performance if the 95% probability interval surrounding their composite score (site difference) fell entirely above 0, as performing

worse than expected if the 95% probability interval surrounding their composite score (site difference) fell entirely below 0, and as expected otherwise. We compared risk-adjusted mortality and complication rates for endpoints in the composite across the three performance groups. We think it is desirable if the three groups have different risk-adjusted performance on each individual metric, as expected.

Assessment of case mix adjustment model

Validity of the proposed measure depends in part on the adequacy of the risk adjustment model to adjust for case mix. As such, our empirical validity testing focused heavily on assessing consistency between the observed data and the underlying assumptions used for statistical analysis. Specifically, we created calibration plots for the overall cohort and for several pre-specified subgroups. Large discrepancies between observed and model-predicted probabilities in any of the plots would suggest that the functional form of the model was misspecified and that estimates of provider performance may be invalid. Methods and results of these analyses are provided in Section 2b4.

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

Associations of component endpoints with 1-year mortality and KCCQ

All 4 non-fatal periprocedural complications in the composite were found to be associated with increased risk of one year mortality. Increased mortality risk was observed in patients with perioperative stroke (adjusted HR 2.10; 95% CI 1.65 to 2.87; p<0.001), major or life-threatening bleeding (adjusted HR 1.92; 95% CI 1.42 to 2.60, p<0.001), modified AKIN Stage III acute kidney injury (adjusted HR 1.81; 95% CI 1.38 to 2.37, p<0.001), and moderate or severe peri-valvular aortic regurgitation (adjusted HR 1.50; 95% CI 1.24 to 1.81; p<0.001).

Similarly, non-fatal periprocedural complications in the composite were also associated with 1-year patient reported health status as assessed by the KCCQ-OS score. Any stroke (adjusted impact on 1-year KCCQ-OS -5.8 points; 95% CI -9.2 to -2.4, p<0.001) and moderate or severe paravalvular regurgitation (adjusted KCCQ-OS impact -2.0 points; 95% CI-3.8 to -0.30, p=0.021) were independently associated with poorer adjusted KCCQ-OS at one year. Modified AKIN Stage III acute kidney injury (adjusted KCCQ-OS impact -3.3 points; 95% CI -6.8 to 0.28, p=0.07) and major or life-threatening bleed (adjusted KCCQ-OS impact 0.4 points; 95%CI -2.0 to 1.2, p=0.619) were not associated with one year KCCQ-OS but were retained in the global rank composite measure, given their strong associations with 1-year mortality.

Associations of component endpoints with categories based on overall composite

Adjusted observed to expected (O/E) ratios of the individual endpoint components according to the 3 levels of site performance are summarized in the table. Sites with better than expected performance on the global rank composite metric showed lower O/E ratios for all components of the global rank composite measure compared with the sites that performed as expected or worse than expected. Similarly, sites with worse than expected performance on the global rank composite demonstrated consistently higher O/E ratios than the other sites. The largest differences favoring the better than expected sites were observed in the incidence of major, life threatening or disabling bleeding and moderate or severe paravalvular leak.

Variables	Observed to Expected Ratios Better than Expected (Sites = 25) (N = 7993)	Observed to Expected Ratios As Expected (Sites = 242) (N = 37473)	Observed to Expected Ratios Worse than Expected (Sites = 34) (N = 7095)
Death	0.71	1.01	1.25

Variables	Observed to Expected Ratios Better than Expected (Sites = 25) (N = 7993)	Observed to Expected Ratios As Expected (Sites = 242) (N = 37473)	Observed to Expected Ratios Worse than Expected (Sites = 34) (N = 7095)
	(2.25% / 3.16%)	(3.21% / 3.17%)	(4.06% / 3.26%)
Stroke	0.73 (1.74% / 2.38%)	1.03 (2.49% / 2.41%)	1.29 (3.16% / 2.44%)
Major or Life Threatening /Disabling Bleed	0.45 (2.84% / 6.30%)	1.02 (6.48% / 6.33%)	2.13 (13.6% / 6.38%)
Acute Kidney Injury or New Dialysis	0.67 (0.83% / 1.23%)	1.12 (1.34% / 1.20%)	1.17 (1.44% / 1.23%)
Moderate or Severe Peri-valvular Regurgitation	0.77 (1.88% / 2.45%)	1.19 (2.71% / 2.28%)	2.00 (4.76% / 2.38%)

Assessment of case mix adjustment model

Results are presented in Section 2b4.

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

Empirical analyses confirm the importance of the selected endpoints by demonstrating their significant associations with one-year mortality and functional status.

The test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance. These results support the validity of the composite measure as a quality measure for TAVR.

2b2. EXCLUSIONS ANALYSIS

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

NA ☐ no exclusions — skip to section 2b4

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

As noted in Section 1.6, the analysis was restricted to hospitals with ≥90% complete non-missing data for 30-day outcome status, baseline KCCQ-12 score, and baseline gait speed. Analyses testing the impact of this exclusion are described in the section Sections 2b6.1-2b6.3 on missing data handling.

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

See Section 2b6.2.

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

As detailed in Section 2b6.3, the exclusion of a high proportion of sites with inadequate data completeness had minimal impact on the classification results of sites that met inclusion criteria for the measure.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

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

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b4.

2b3.1. What method of controlling for differences in case mix is used? *(check all that apply)*

☐ **Endorsed (or submitted) as individual performance measures**

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

The global ranking endpoint is an ordinal categorical variable having 6 levels where category 1 represents the worst possible outcome (death) and category 6 represents the best possible outcome (alive and free of major complications). Variation in the distribution of outcome categories across hospitals was modeled by a hierarchical proportional odds model with hospital-specific random effects (intercepts) [1]. The proportional odds methodology is an extension of binary logistic regression for multiple ordered outcome categories [2]. We implemented this methodology within a multilevel hierarchical modeling framework in order to estimate hospital-specific summary metrics while also adjusting for case mix and simultaneously accounting for the clustered data structure [3-5]. Because extreme outcomes can result from random statistical fluctuations, we estimated intercepts using an empirical Bayes (aka reliability-adjusted [6]) methodology which shrinks noisy extreme estimates toward the null value of zero [7]. To further account for uncertainty, we calculated a 95% empirical Bayes probability interval around each hospital's intercept estimate.

Details of Statistical Model

The mathematical form of the model is

$$\log\frac{\Pr(Y_{hi} \le k)}{\Pr(Y_{hi} > k)} = \alpha_k + \sum_{j=1}^{x} \beta_j x_{hij} + e_h, \quad k = 1, \ldots, 5$$

$$e_h \sim \text{normal}(\text{mean} = 0, \text{variance} = \sigma^2)$$

where Y_{hi} is the outcome category (1=death, 2=stroke, etc.) of the i-th patient from the h-th hospital, x_{hij} is a numerical representation of the j-th baseline covariate of the i-th patient from the h-th hospital, and the quantities $\alpha_1 \ldots, \alpha_5, \beta_1, \beta_2, \ldots, \beta_Q$ and σ^2 are unknown parameters (fixed effects) to be estimated from the data. The random effect e_h captures the h-th hospital's tendency to have outcomes in lower ranking categories with larger numbers implying worse outcomes. Fixed effect parameter estimates $\hat{\alpha}_k, \hat{\beta}_j$, and $\hat{\sigma}^2$

were calculated via maximum likelihood and random effect estimates $\hat{e}_h$ were calculated using an empirical Bayes shrinkage estimator as implemented in SAS PROC GLIMMIX.

Calculation of Site Difference Metric

Information about a site's relative performance according to the model is encapsulated in the hospital's intercept parameter, but intercepts do not have a simple or intuitive interpretations for clinicians. To address this, we transformed intercept estimates into an interpretable metric that resembles the win-ratio [8] and net benefit [9] approaches that have been proposed for assessing ranked outcomes in clinical trials based on the proportions of winners and losers in paired analyses. In a conventional win ratio analysis, the win ratio is estimated by forming all possible pairs of 1 patient from the treatment group and 1 patient from the control group and then dividing the proportion of pairs for which the treated patient has a better outcome by the proportion of pairs for which the untreated patient has a better outcome. The calculation of the net benefit metric is identical except that it subtracts losing from winning pairs instead of dividing them. The TVT site difference metric is conceptually similar to the net benefit approach but instead of two treatments we compare each hospital to a hypothetical average "reference" hospital. Another difference is that we use a parametric instead of nonparametric estimator and use the parametric approach to adjust for case mix.

Conceptually, the site metric is calculated by pairing each patient treated by the TAVR hospital of interest with a hypothetical patient having identical risk factors who is treated by a hypothetical average hospital (intercept = 0). The "site difference" metric is calculated as the model-predicted proportion of winning pairs minus losing pairs where "winning pair" means that the hospital of interest's patient had a better outcome compared to the reference hospital and "losing pair" means that the reference hospital's patient had a worse outcome compared to the reference hospital. A site difference greater than zero implies that the hospital of interest's outcomes are better than expected in light of its case mix and a site difference less than zero implies that the hospital of interest's outcomes are worse than expected in light of its case mix.

Details of Site Difference Calculation

For a generic TAVR patient, let Y be the patient's outcome category and let $\pi_k(\boldsymbol{x}; e)$ be the model-predicted probability of the k-th outcome category for a patient with baseline covariate vector equal to $\boldsymbol{x}$ who is treated at a hospital with an intercept parameter equal to e, that is:

$$\pi_k(\boldsymbol{x}; e) = \Pr(Y = k | \text{covariates} = \boldsymbol{x}, \text{intercept} = e).$$

The expression $\pi_k(\boldsymbol{x}; e)$ can be calculated for any choice of $\boldsymbol{x}$ and e based on output from the fitted model. Let Y_{hi} be the outcome of the i-th patient from the h-th hospital, let $\boldsymbol{x}_{hi} = (x_{hi1}, \ldots, x_{hiQ})$ be this patient's set of baseline covariates in the model, and let Y_{hi}^{ref} be the outcome of a hypothetical patient with identical covariates who is treated by a reference hospital with intercept equal to 0. Let W_{hi} denote the probability of a win in the sense that $Y_{hi} > Y_{hi}^{\text{ref}}$ and let L_{hi} denote the probability of a loss in the sense that $Y_{hi} < Y_{hi}^{\text{ref}}$:

$$W_{hi} = \sum_{k=1}^{5} \pi_k(\boldsymbol{x}_{hi}; 0)[\pi_{k+1}(\boldsymbol{x}_{hi}; e_h) + \cdots + \pi_6(\boldsymbol{x}_{hi}; e_h)]$$

$$L_{hi} = \sum_{k=1}^{5} \pi_k(\boldsymbol{x}_{hi}; e_h)[\pi_{k+1}(\boldsymbol{x}_{hi}; 0) + \cdots + \pi_6(\boldsymbol{x}_{hi}; 0)]$$

The site difference metric for the h-th hospital is calculated as

$$\text{site-difference}_h = \left(\sum_{i=1}^{n_h} W_{hi} - \sum_{i=1}^{n_h} L_{hi} \right) \Big/ n_h$$

where n_h is the number of eligible patients from the h-th hospital.

References

1. Agresti A, Booth JG, Hobert JP, Caffo B. 2. Random-Effects Modeling of Categorical Response Data. Sociological Methodology. 2000 Aug;30(1):27-80.
2. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015 Aug 14.
3. Daniels MJ, Gatsonis C. Hierarchical polytomous regression models with applications to health services research. Statistics in Medicine. 1997 Oct 30;16(20):2311-25.
4. Daniels MJ, Gatsonis C. Hierarchical generalized linear models in the analysis of variations in health care utilization. Journal of the American Statistical Association. 1999 Mar 1;94(445):29-42.
5. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, Howard G, Peterson ED, Rathore SS, Smith Jr SC, Spertus JA. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. 2006 Jan 24;113(3):456-62.
6. Dimick JB, Ghaferi AA, Osborne NH, Ko CY, Hall BL. Reliability adjustment for reporting hospital outcomes with surgery. Annals of surgery. 2012 Apr 1;255(4):703-7.
7. Ten Have TR, Localio AR. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. Biometrics. 1999 Dec;55(4):1022-9.
8. Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012 Jan;33(2):176-82.
9. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. JAMA oncology. 2016 Jul 1;2(7):901-5.

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

2b3.3a. Describe the conceptual/clinical *and* statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion;** for example, are social risk factors added after all clinical factors?
Covariates for case mix adjustment were pre-selected to be the same as the NQF-endorsed (3534) TVT TAVR 30-day mortality measure. We favored a non-parsimonious model in order to reduce the risk of confounding by unmeasured case mix factors and better satisfy causal inference assumptions. In general, our goal was to adjust for all potential confounders of the observed associations between site and composite outcomes. In general, we only adjusted for pre-treatment factors and avoided adjusting for aspects of the patient's treatment.

Age	Prior peripheral artery disease	# prior cardiac operations
BSA	Current/recent smoker	Prior aortic procedure
Sex	Diabetes	Prior other valve procedure

Age	Prior peripheral artery disease	# prior cardiac operations
Race/ethnicity	NYHA Class	Aortic etiology
eGFR	Atrial fibrillation/flutter	Valve morphology
Dialysis	Conduction defect	Aortic insufficiency
Ejection fraction	Chronic lung disease	Mitral insufficiency
Hemoglobin	Home oxygen	Tricuspid insufficiency
Platelet count	Hostile chest	Acuity status
Procedure date	Porcelain aorta	Cardiogenic shock
LMD ≥ 50%	Access site	Cardiac arrest w/in 24 hours
Proximal LAD ≥ 70%	Pacemaker	Pre-procedure inotropes
Prior MI	Previous ICD	Mechanical assist device
Endocarditis	Prior PCI	Carotid stenosis
Gait speed	Prior CABG	Prior TIA/stroke
Baseline KCCQ-12	*	*

*cell intentionally left blank

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

- ☐ **Published literature**
- ☒ **Internal data analysis**
- ☐ **Other (please describe)**

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

Covariates were pre-selected and were therefore retained in the model regardless of their apparent statistical significance. The statistical significance of a covariate's association with outcomes is partly a reflection of sample size. A covariate with a non-significant p-value in the development sample could have a significant association in a future production run of the composite measure. Estimated covariate effects in the development sample are summarized in the following table.

Select Risk Factors	Odds Ratio	P-value
Age (per year)	0.99 (0.98, 1.00)	0.054
Age (per year when > 75)	1.02 (1.01, 1.03)	<.001
BSA (per m^2 increase) for male	0.53 (0.44, 0.63)	<.001
BSA (per m^2 increase) for female	0.40 (0.34, 0.48)	<.001
LVEF (per 1%)	1.00 (1.00, 1.00)	0.465
Hemoglobin (per g/dL)	0.98 (0.97, 1.00)	0.062
Platelet Count (per unit)	1.00 (1.00, 1.00)	<.001
Platelet count (per unit when >200k)	1.00 (1.00, 1.00)	<.001
Procedure Date	1.00 (1.00, 1.00)	<.001

Select Risk Factors	Odds Ratio	P-value
GFR (per unit)	0.99 (0.99, 0.99)	<.001
Current Dialysis	0.81 (0.70, 0.95)	0.009
Female	1.62 (1.04, 2.51)	0.033
Non-White or Hispanic	0.96 (0.87, 1.06)	0.461
Left Main >=50%	1.10 (1.00, 1.22)	0.041
Proximal LAD	1.20 (1.11, 1.30)	<.001
Prior MI	1.03 (0.96, 1.10)	0.401
Endocarditis	1.10 (0.84, 1.44)	0.480
Prior TIA w/o Stroke, or Prior Stroke	1.03 (0.97, 1.10)	0.375
Carotid Stenosis One or Both	1.04 (0.98, 1.11)	0.229
Prior PAD	1.20 (1.14, 1.28)	<.001
Smoker	1.09 (0.98, 1.21)	0.123
Diabetes	0.98 (0.92, 1.03)	0.396
Afib/Flutter	1.07 (1.01, 1.13)	0.013
Conduction Defect	1.04 (0.98, 1.10)	0.173
CLD (Severe)	1.22 (1.12, 1.32)	<.001
Home Oxygen	1.12 (1.02, 1.22)	0.012
Hostile Chest	1.02 (0.92, 1.14)	0.705
Porcelain Aorta	1.15 (1.02, 1.30)	0.026
Non-Femoral Access	1.71 (1.57, 1.86)	<.001
Pacemaker	0.85 (0.78, 0.91)	<.001
Previous ICD	1.02 (0.89, 1.18)	0.754
Prior PCI	1.03 (0.98, 1.10)	0.239
Prior CABG	0.87 (0.77, 0.98)	0.026
Prior Cardiac Operations (1 vs. 0	0.88 (0.79, 0.99)	0.036
Prior Cardiac Operations (2+ vs. 0)	0.99 (0.81, 1.19)	0.882
Prior Aortic Procedure	0.94 (0.86, 1.03)	0.182
Prior Non-Aortic Procedure	1.16 (0.97, 1.38)	0.107
Aortic Etiology (Degenerative vs. Other)	0.96 (0.84, 1.09)	0.520
Valve Morphology (Tricuspid)	1.00 (0.91, 1.10)	0.974
Aortic Insufficiency (Moderate/Severe)	0.94 (0.88, 1.00)	0.060
Mitral Insufficiency (Moderate/Severe)	1.03 (0.97, 1.10)	0.333
Tricuspid Insufficiency (Moderate/Severe)	1.09 (1.02, 1.16)	0.007

Select Risk Factors	Odds Ratio	P-value
Urgent vs. Elective	1.20 (1.08, 1.33)	<.001
Shock/Inotrope/DeviceAssist vs. Elective	1.63 (1.41, 1.88)	<.001
Emergency/Salvage/CardiacArrest vs. Elective	2.72 (2.02, 3.65)	<.001
Unable to Walk	1.12 (1.02, 1.23)	0.022
Gait Speed (per m/s)	0.72 (0.64, 0.82)	<.001
KCCQ Overall Score (per unit)	1.00 (1.00, 1.00)	<.001

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

For the purposes of this NQF submission and the question on whether social risk factors should be included in risk adjustment, we modeled this composite to match the covariates used in NQF #3534, 30 Day All-cause Risk Standardized Mortality Odds Ratio following Transcatheter Aortic Valve Replacement (TAVR). During testing of that measure, we examined variables for black race, other non-white race, Hispanic ethnicity, and participation in Medicaid and whether any of these variables had any statistically significant associations with 30-day mortality after adjusting for other factors in the hierarchical model. For each variable in each time period, the 95% confidence interval around the odds ratio overlapped with the null value of 1.0, which implies that there was no statistically significant association. As a result, we did not collect or analyze any variables that directly measure social risk in this composite beyond what is already included. The model covariates that are presumed to be directly or indirectly associated with aspects of social risk include age, sex, race, and ethnicity, among others.

2b3.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.

If stratified, skip to 2b3.9

We performed a series of analyses to assess consistency between the observed data and the underlying assumptions used for statistical analysis. Our approach emphasized graphical displays as opposed to hypothesis testing because all models are only approximations and p-values are dependent on sample size. Our goal was to assess whether the net effect of any modeling errors (e.g., failure of the model's underlying proportional odds assumption) would significantly impact agreement between observed and predicted probabilities. For each patient, the model uses the patient's risk factors to predict the patient's probability of having an outcome in one of the k worst global ranking categories, k=1,2,...,5, where k=1 means the probability of death, k=2 means the probability of death or stroke, etc. For each of these 5 probabilities, we assessed calibration by plotting observed versus average model-predicted proportions across 10 equally sized groups of patients based on a ranking of their predicted probabilities. We created calibration plots for the overall cohort and for several pre-specified subgroups. Large discrepancies between observed and model-predicted probabilities in any of the plots described above would suggest that the functional form of the model was misspecified.

To assess calibration at the level of individual hospitals, we grouped patients by hospital and compared observed versus model-predicted proportions within each hospital. To identify hospitals with large discrepancies, we calculated a 99% confidence interval around each hospital's observed proportion and counted the number of hospitals for which the 99% confidence interval was not overlapping the predicted probability. Intuitively, one might expect approximately ~1% of the 301 hospitals (i.e., 3 hospitals) to have a non-overlapping confidence interval if the fit to the data was excellent.

In addition to assessing calibration, we also estimated the C-index (i.e., discrimination) for predicting each dichotomization of the global ranking endpoint. The C-index quantifies the ability of a classification algorithm to separate the target population into a group of patients who will have the endpoint of interest and a group of patients who will not have the endpoint of interest. A low C-index does not imply that the model is misspecified or that hospital comparisons will be biased. Nonetheless, the C-index is widely reported and was, therefore, calculated for the sake of completeness. The C-index was not adjusted for optimism because our goal was to fit the data at hand rather than to use model coefficients to predict future outcomes.

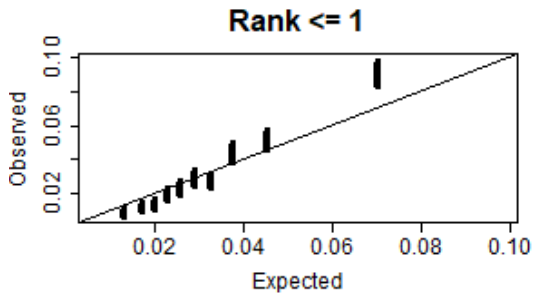
2b3.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

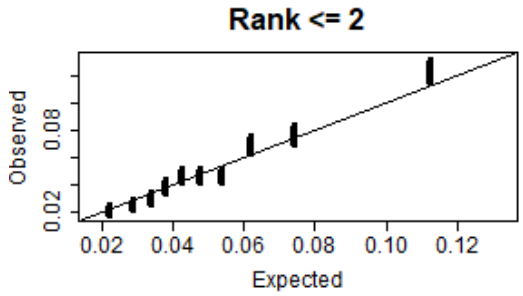
Table: C-statistic for predicting an outcome in one of the worst ranking categories

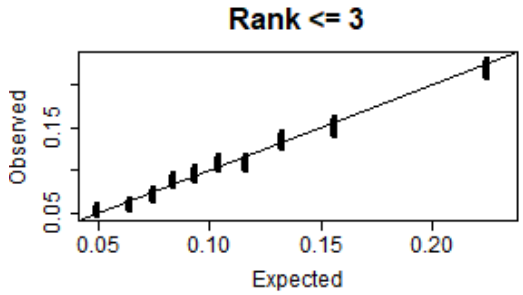
Rank ≤ 1	Rank ≤ 2	Rank ≤ 3	Rank ≤ 4	Rank ≤ 5
0.70	0.65	0.63	0.64	0.63

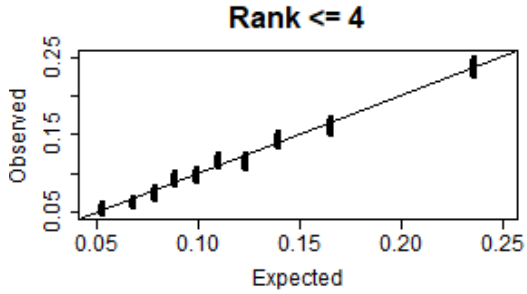
2b3.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*): N/A

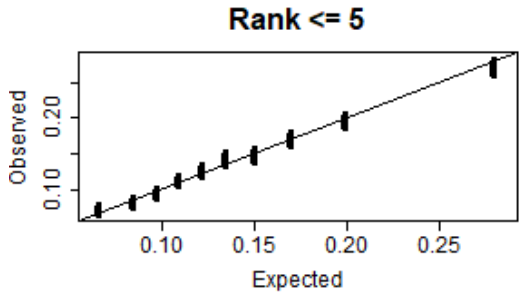
2b3.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:
Overall Calibration


Rank <= 1
Observed
0.02
0.06
0.10
0.02
0.04
0.06
0.08
0.10
Expected

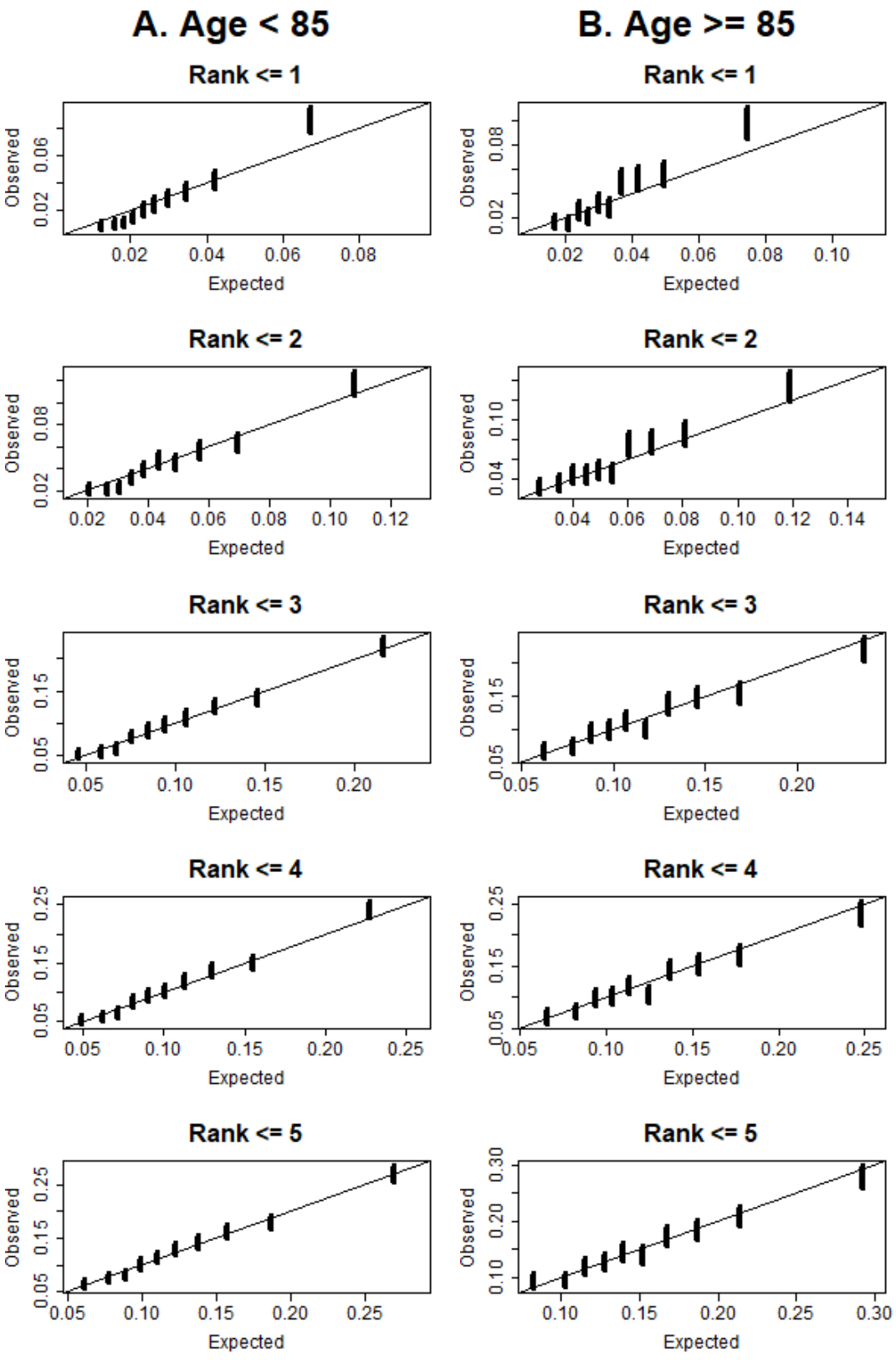

Rank <= 2
Observed
0.02
0.08
0.02
0.04
0.06
0.08
0.10
0.12
Expected

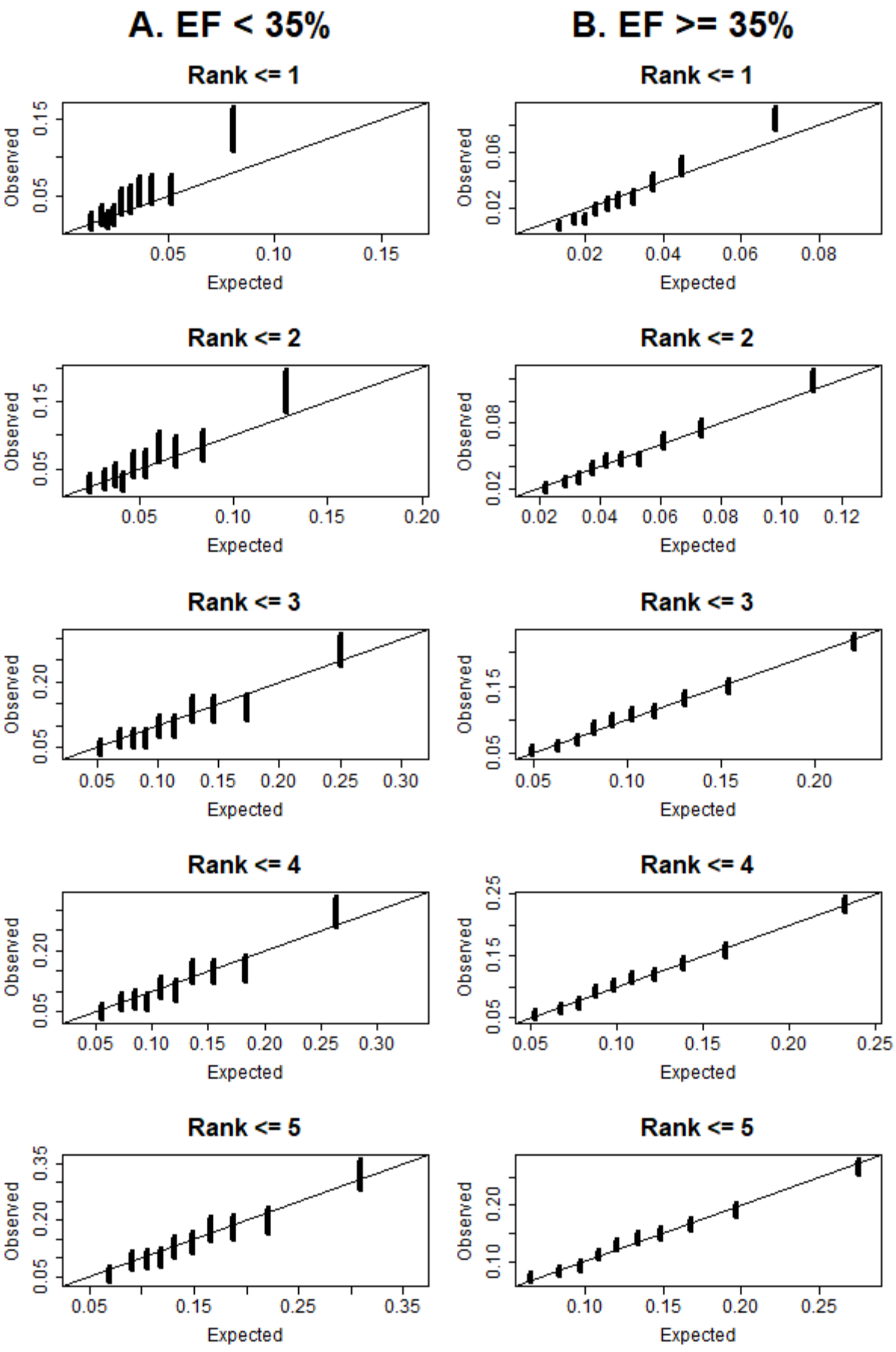

Rank <= 3
Observed
0.05
0.15
0.05
0.10
0.15
0.20
Expected

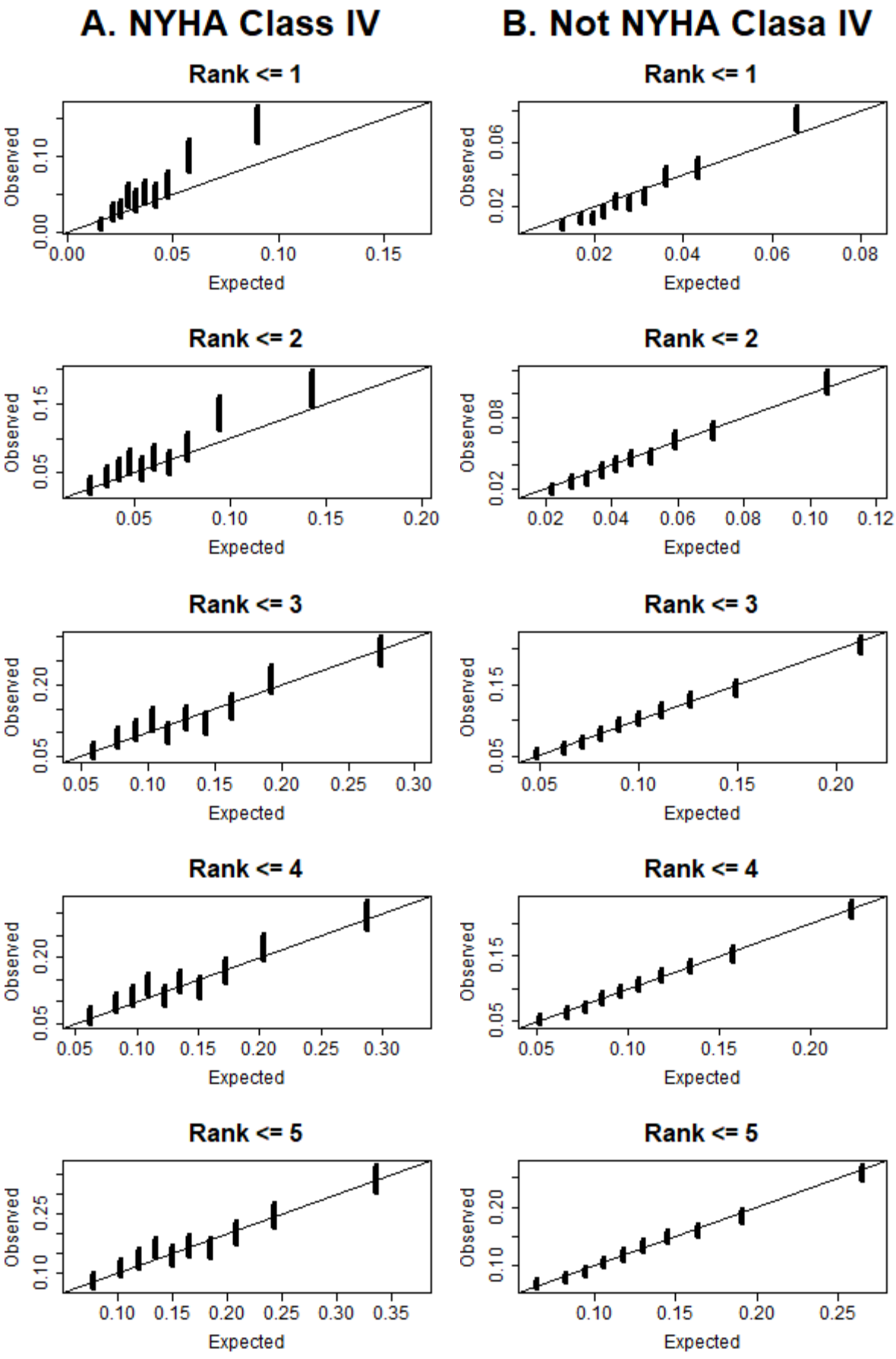

Rank <= 4
Observed
0.05
0.15
0.25
0.05
0.10
0.15
0.20
0.25
Expected

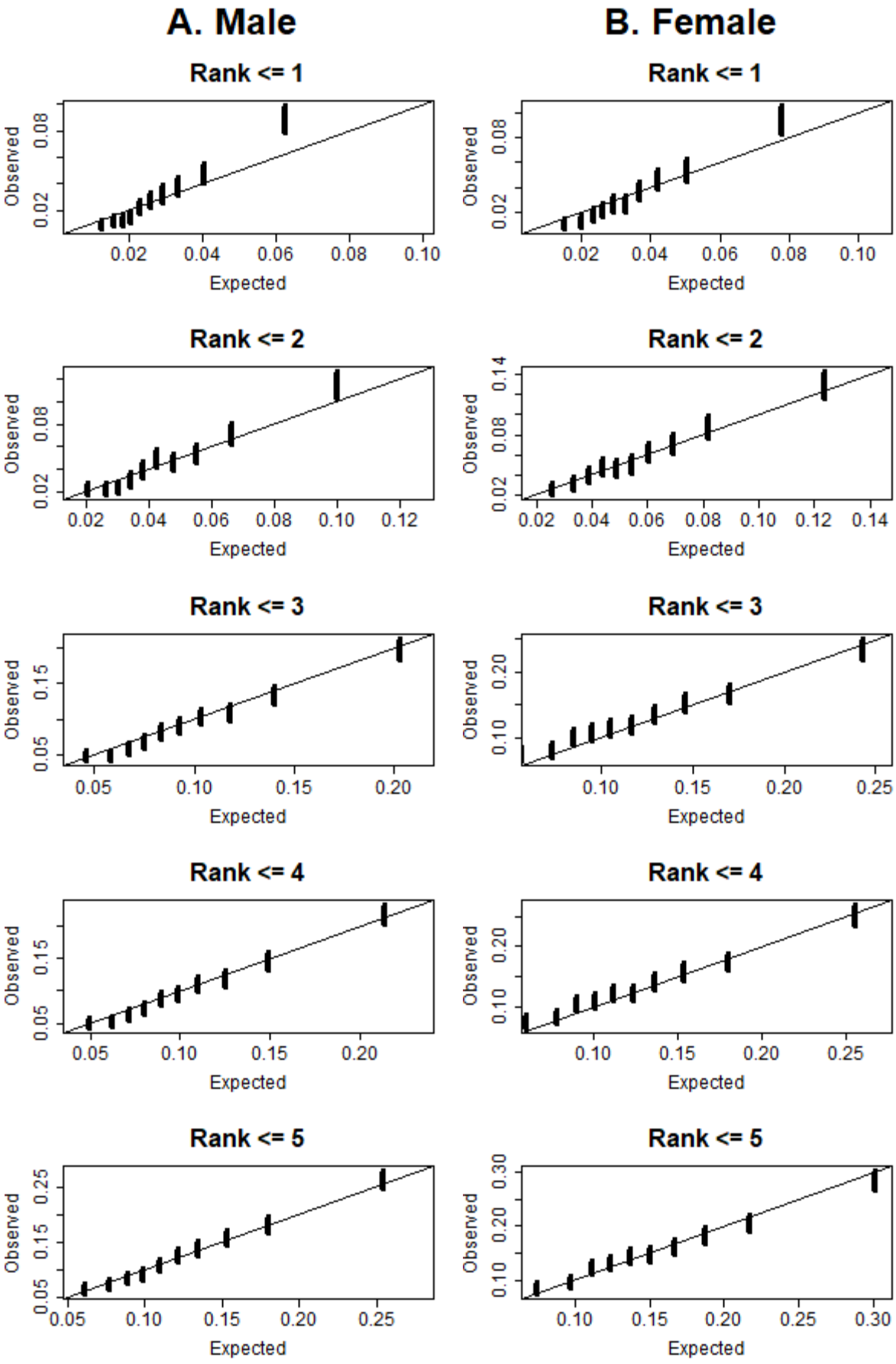

Rank <= 5
Observed
0.10
0.20
0.10
0.15
0.20
0.25
Expected

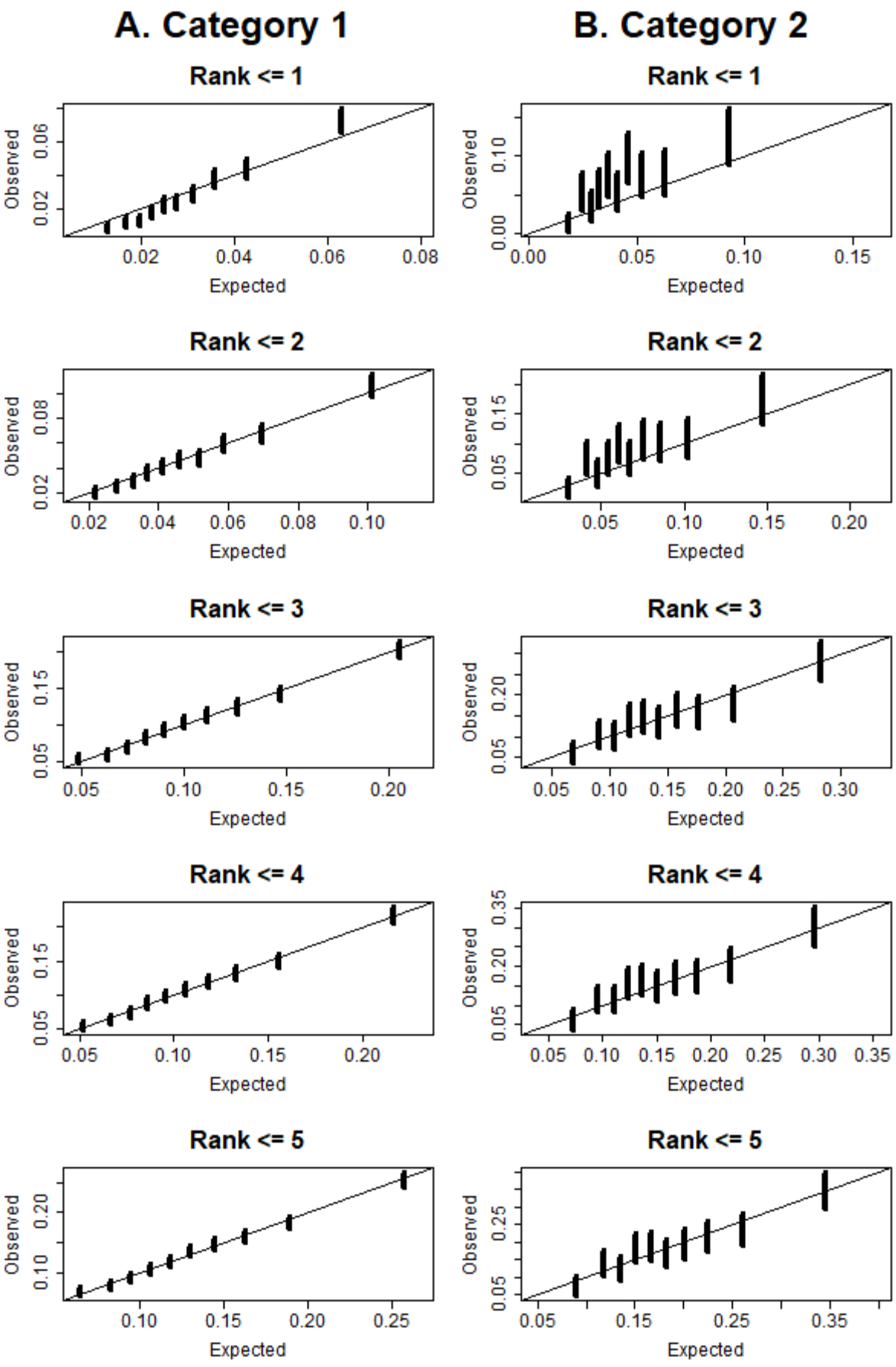
Calibration within pre-specified subgroups



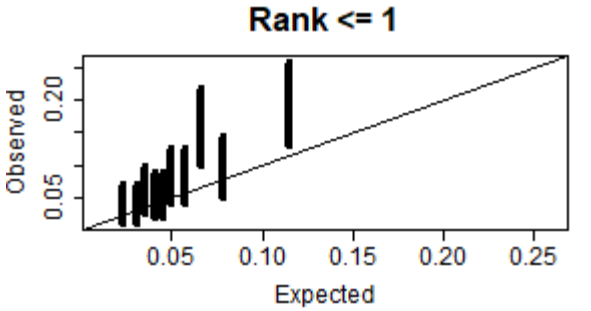

A. EF < 35%
B. EF >= 35%
Rank <= 1
Rank <= 2
Rank <= 3
Rank <= 4
Rank <= 5
Observed
Expected

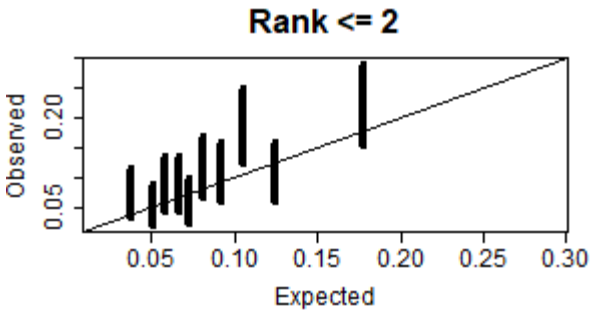

A. NYHA Class IV
B. Not NYHA Clasa IV
Rank <= 1
Rank <= 2
Rank <= 3
Rank <= 4
Rank <= 5
Observed
Expected

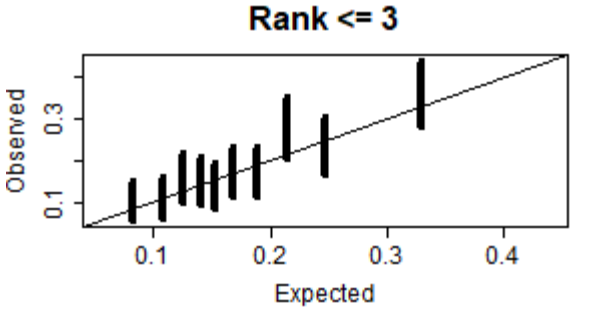

A. Male
B. Female
Rank <= 1
Rank <= 2
Rank <= 3
Rank <= 4
Rank <= 5
Observed
Expected

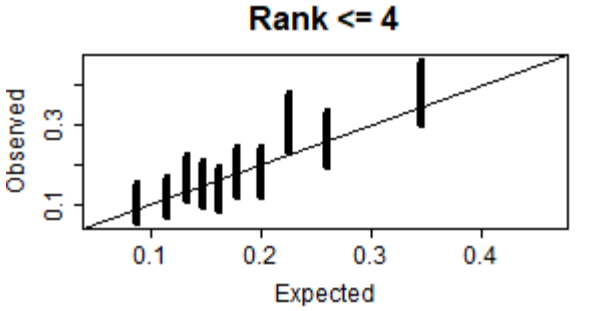

A. Category 1
B. Category 2
Rank <= 1
Rank <= 2
Rank <= 3
Rank <= 4
Rank <= 5
Observed
Expected

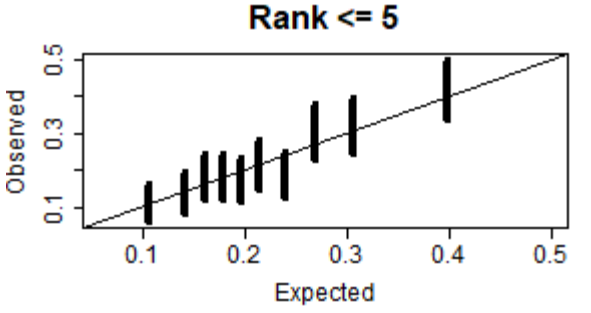
C. Category 3



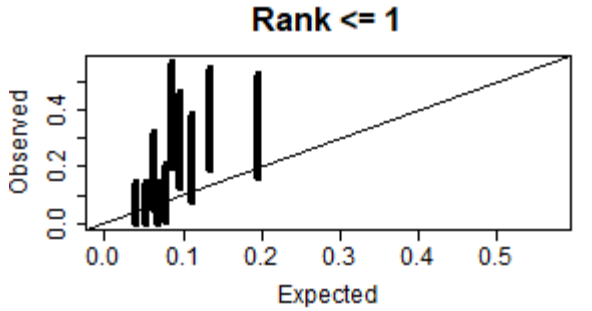


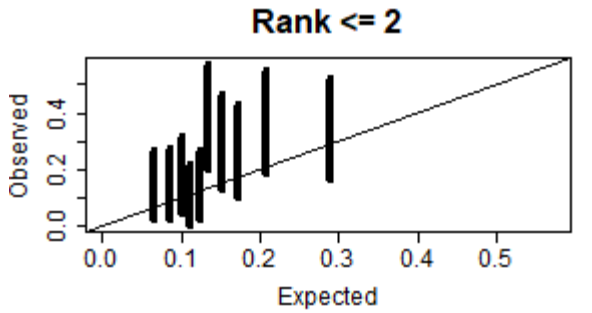


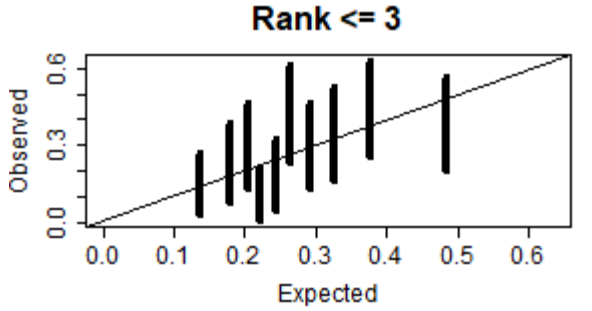


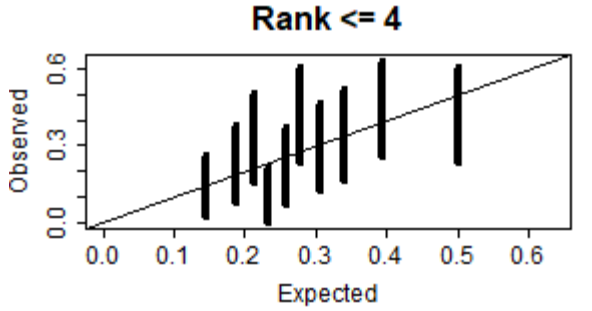


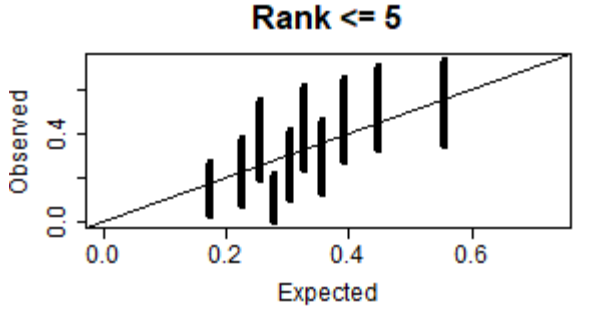
D. Category 4

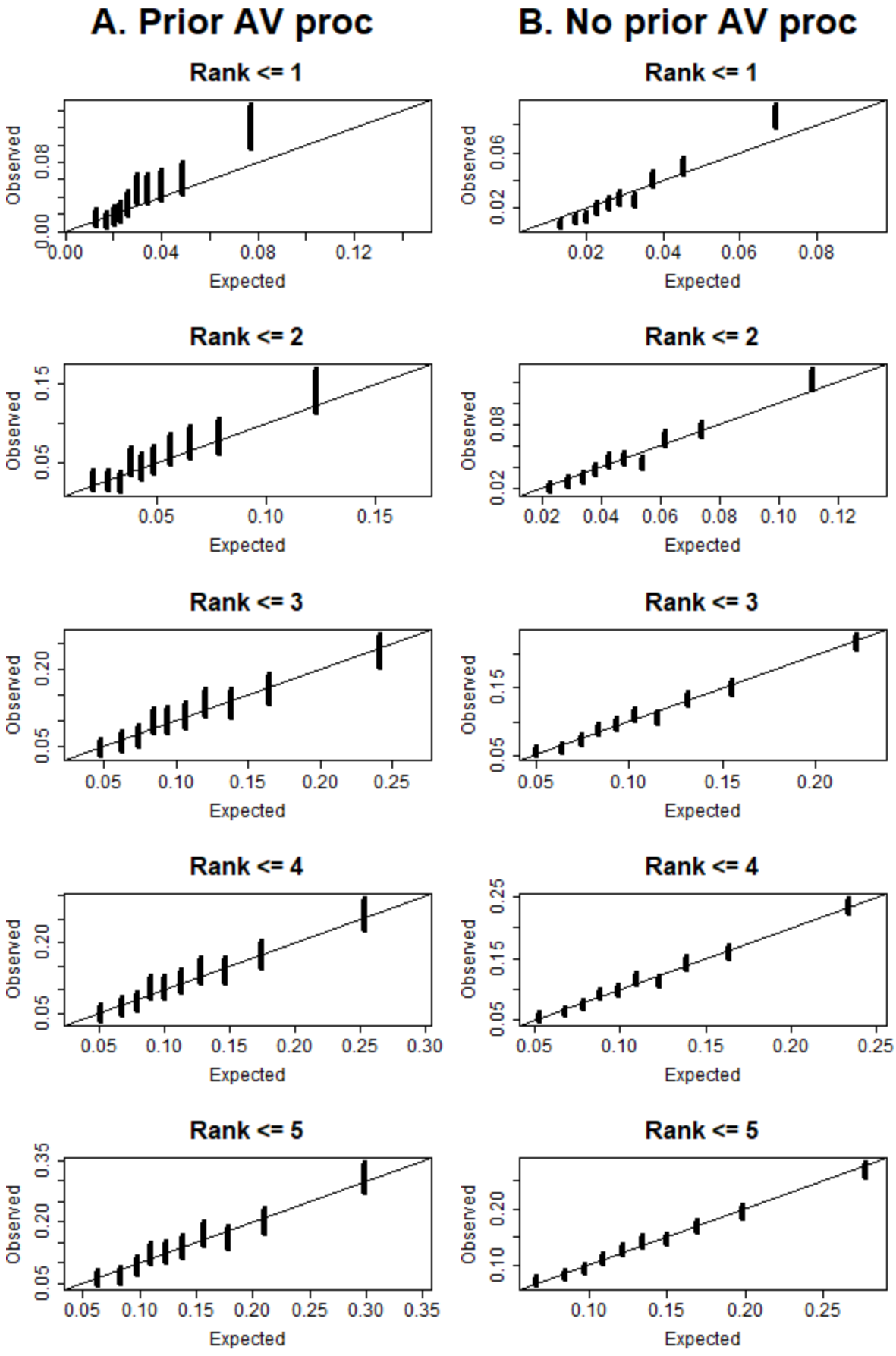












Hospital-specific goodness of fit testing results

To identify hospitals with large discrepancies, we calculated a 99% confidence interval around each hospital's observed proportion and counted the number of hospitals for which the 99% confidence interval was not overlapping the predicted probability. Intuitively, one might expect approximately ~1% of the 301 hospitals (i.e., 3 hospitals) to have a non-overlapping confidence interval if the fit to the data was excellent. As shown in the table, results were generally consistent with the hypothesis of adequate model calibration.

Rank	Model Under-Estimates Observed Events: Number of Hospitals	Model Over-Estimates Observed Events: Number of Hospitals
Rank ≤ 1	3	1
Rank ≤ 2	1	0
Rank ≤ 3	5	2
Rank ≤ 4	3	0
Rank ≤ 5	1	0

2b3.9. Results of Risk Stratification Analysis: N/A

2b3.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)
Results demonstrate that the model adopted for case mix adjustment provides an adequate fit and is well suited to adjust for case mix.

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

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

Note: *Applies to the composite performance measure.*

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

The degree of uncertainty surrounding a hospital's composite measure estimate is indicated by calculating a 95% empirical Bayes probability interval around the hospital's site difference metric. Point estimates and probability intervals are reported along with a comparison to the null value of zero. In addition, the composite measure result is converted into categories. A hospital is categorized as having better than expected performance if the lower limit of the hospital's 95% probability interval is greater than 0. Otherwise, a hospital is classified as having worse than expected performance if the upper limit of the hospital's 95% probability interval is less than 0. Otherwise, the hospital is classified as performing as expected.

The composite measure's ability to identify statistically and clinically meaningful differences in performance was assessed by tabulating the number of hospital's classified as performing better, worse, or same as expected based on the site's 95% probability interval.

To assess the clinical or practical significance of the observed differences, we also plotted a histogram of the site-specific performance estimates. These estimates incorporate a reliability adjustment (via hierarchical modeling; shrinkage estimation) which causes the estimates to be slightly biased toward zero. As a result, the histogram of site-specific shrinkage estimates provides a conservative representation of the true magnitude of between-site differences. The actual true signal variation between hospitals is likely to be larger than shown in the histogram.

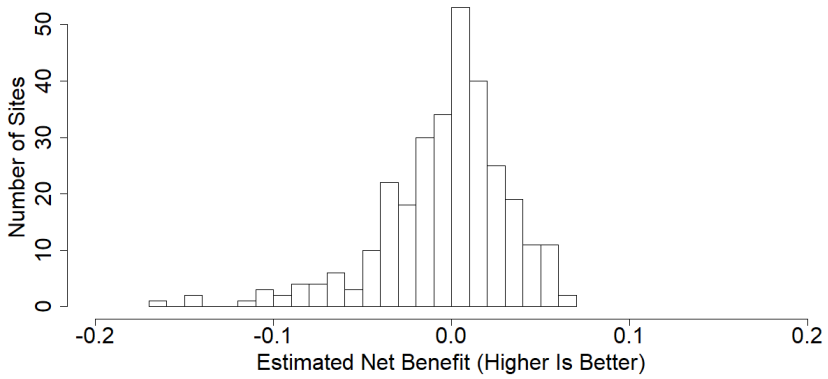
2b4.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly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

Number of Hospitals By Statistical Categorization Based on 95% Interval

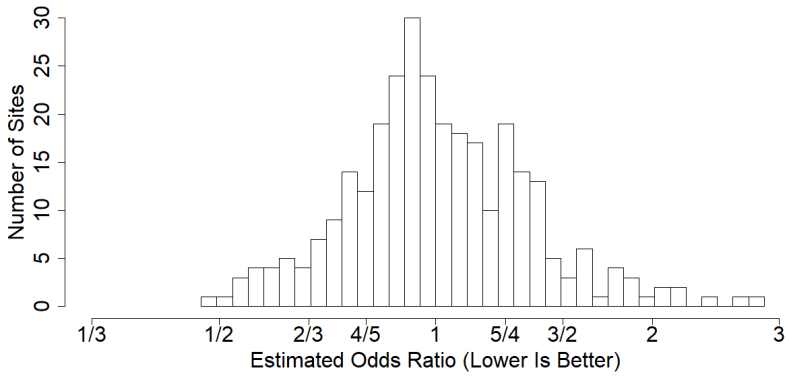
Better Than Expected	As Expected	Worse Than Expected
25 / 301 (8%)	242 / 301 (80%)	34 / 301 (11%)

Note: Based on whether 95% interval around the site metric falls entirely below above zero (better), entirely below zero (worse), or overlaps zero (as expected).

Figure: Histogram of site-specific composite score estimates.



As detailed in Section 2b3.1.1, the site difference is derived from each hospital's intercept parameter (along with other parameters) in the hierarchical model. The site difference is greater than zero if and only if the intercept parameter is less than zero. To provide an alternative perspective for assessing between-site differences, we converted intercept parameters into odds ratios and plotted the odds ratios. The odds ratio represents the ratio of a patient's odds of having a poor outcome if treated by the hospital of interest compared to the odds of having a poor outcome if treated by an average hospital. Estimated odds ratio range from less than 1/2 to almost 3 representing an approximately 6-fold variation in the odds of a poor outcome across TAVR hospitals.



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

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

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

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

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

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

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

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

---

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

Note: *Applies to the overall composite measure.*

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

The analysis was restricted to hospitals with ≥90% complete non-missing data for 30-day outcome status, baseline KCCQ-12 score, and baseline gait speed. To explore how included and excluded TVT centers might differ, we compared site- and patient-level characteristics between included and excluded sites. We calculated median and interquartile range for continuous variables and percentages for categorical variables.

To assess whether the exclusion of sites with <90% data completeness impacted classification results of included sites, we performed a sensitivity analysis in which KCCQ-12 score and gait speed were removed from the risk adjustment model and sites with ≥90% missing data for these variables were included.

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

Table: Characteristics of hospitals that were included versus excluded on the basis of inadequate data completeness.

Hospital Characteristics	Overall Sites = 556	Excluded Sites = 255	Included Sites = 301
Annual TAVR volume, median (IQR)	61.5 (42.6, 99.5)	67.4 (44.0, 115)	58.7 (41.7, 90.9)
Teaching hospital, n (%)	323 (58.1)	154 (60.4)	169 (56.1)
Region: South	204 (36.7)	91 (35.7)	113 (37.5)
Region: Midwest	130 (23.4)	48 (18.8)	82 (27.2)
Region: West	121 (21.8)	57 (22.4)	64 (21.3)
Region: Northeast	98 (17.6)	57 (22.4)	41 (13.6)
Region: Missing	3 (0.5)	2 (0.8)	1 (0.3)
Location: Rural	52 (9.4)	19 (7.5)	33 (11.0)
Location: Suburban	148 (26.6)	71 (27.8)	77 (25.6)
Location: Urban	356 (64.0)	165 (64.7)	191 (63.5)
No. of Patient Beds, median (IQR)	456 (332, 636)	500 (342, 689)	438 (317, 608)

Table: Characteristics of patients at sites that were included versus excluded on the basis of inadequate data completeness.

Characteristic	Excluded Sites n=61,560)	Included Sites (n=52,561)	Standardized Difference
Age, years	82.0 (76.0, 87.0)	82 (76, 87)	0.029
Male sex, %	54.2	53.5	0.015
Diabetes mellitus, %	37.9	38.4	0.012
Current smoker, %	5.9	5.8	-0.016
Currently on dialysis, %	4.1	3.5	-0.039
GFR, ml/min/1.73 m2	63 (48, 79)	63 (48, 78)	-0.007

Characteristic	Excluded Sites n=61,560)	Included Sites (n=52,561)	Standardized Difference
LVEF, %	58 (48, 60)	58 (50, 60)	0.033
Prior MI, %	22.9	22.8	0.006
Prior pacemaker, %	13.8	14.2	0.012
Prior PCI, %	33.5	34.3	0.028
Prior CABG, %	23.4	23.4	-0.002
Prior aortic valve procedure, %	12.2	11.9	-0.01
Prior non-aortic valve procedure, %	2.5	2.2	-0.021
NYHA Class IV, %	14.8	14.8	-0.003
Atrial fibrillation/flutter, %	38.5	39.2	0.011
Conduction defect, %	36.1	38.0	0.036
Prior stroke or TIA, %	17.6	18.5	0.020
Carotid stenosis, %	19.2	21.8	0.061
Peripheral arterial disease, %	28.6	28.2	-0.010
Severe chronic lung disease, %	11.4	11.8	0.012
Home oxygen, %	9.2	10.2	0.03
Hostile chest, %	7.5	7.0	-0.019
Porcelain aorta, %	3.9	4.0	0.002
Non-femoral access, %	8.5	7.6	0.034
Acuity, elective	87.0	90.8	0.126
KCCQ-OS	42.4 (25.0, 62.5)	42.7 (26.0, 62.5)	0.017
5MWT, seconds	7.3 (6.0, 9.7)	7.3 (6.0, 9.7)	0.009

Sensitivity analyses showed that removal of the 90% completeness exclusion for KCCQ-OS and gait speed, thereby allowing for a larger number of sites (444 sites), resulted in nearly identical proportions of classification into the outcome groups and nearly no re-classification within the original 301 site cohort.

Analysis	Better Than Expected	As Expected	Worse Than Expected
Original Analysis Above	25 / 301 (8%)	242 / 301 (80%)	34 / 301 (11%)
Sensitivity Analysis	36/444 (8%)	353/444 (80%)	55/444 (12%)

Reclassification results for 301 hospitals appearing in both the original analysis and sensitivity analysis

Sensitivity Analysis	Original Analysis Better Than Expected	Original Analysis As Expected	Original Analysis Worse Than Expected
Better Than Expected	24	1	0
As Expected	1	241	1
Worse Than Expected	0	0	33

1 site gets reclassified from "Better Than Expected" to "As Expected".

1 site gets reclassified from "As Expected" to "Better Than Expected".

1 site gets reclassified from "Worse Than Expected" to "As Expected".

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

Excluded sites were generally very similar to sites included in the analysis. Excluded sites had a numerically higher mean annualized volume but similar geographic distribution by region, urban versus rural setting and teaching versus non-teaching. Patients at included versus excluded sites were similar. The exclusion of a high proportion of sites with inadequate data completeness had minimal impact on the classification results of sites that met inclusion criteria for the measure.

2c. EMPIRICAL ANALYSIS TO SUPPORT COMPOSITE CONSTRUCTION APPROACH

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

2d1. Empirical analysis demonstrating that the component measures fit the quality construct, add value to the overall composite, and achieve the object of parsimony to the extent possible.

2d1.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used;* ***if no empirical analysis****, provide justification)*

Note: This information is repeated from Sections 2b1.2-2b1.4.

The clinical importance of individual short-term complications was confirmed empirically by assessing their associations with one year mortality and the one-year Kansas City Cardiomyopathy Questionnaire Overall Score (KCCQ-OS) in an earlier TVT TAVR cohort that had been linked with Medicare data in order to follow patients longitudinally. We began by selecting patients aged 65 or older who underwent transfemoral TAVR between January 1, 2015 and January 31, 2016 and were alive 30 days post TAVR, had records linked to Medicare, and had non-missing 30 day data (N=12,607). All records had at least 1 year of follow-up data. Cox proportional hazards modeling was used to evaluate associations of stroke, major or life-threatening bleeding, any major vascular event, new pacemaker, modified AKIN Stage III acute kidney injury, and moderate or severe perivalvular regurgitation with one year mortality. To evaluate associations between non-fatal periprocedural complications and 1 year KCCQ-OS, we included records with non-missing KCCQ-OS (N=10,883) and used inverse probability weighting to adjust for differences between patients who had missing versus non-missing KCCQ-OS. Linear regression models were used to estimate the average change in KCCQ-OS associated with stroke, major or life-threatening bleeding, major vascular event, new pacemaker, modified AKIN Stage III acute

kidney injury, and moderate or severe perivalvular aortic regurgitation. All complications were included simultaneously in both the mortality and KCCQ-OS models to assess each complication, independent of other complications. In addition to complications, covariates in the model included: Age, Sex, Body Surface Area, LVEF, Hemoglobin, Platelet count, GFR, Dialysis, Race White, Hispanic or Latino Ethnicity, Left Main Stenosis>=50%, Proximal LAD >=70%, Prior MI, Endocarditis, Prior Stroke/TIA, Carotid Stenosis, Prior PAD, Smoker, Diabetes, NYHA, Atrial Fibrillation, Conduction Defect, Chronic Lung Disease, Home Oxygen, Hostile Chest, Porcelain Aorta, Non-Femoral Access Site, Pacemaker, Previous ICD, Prior PCI, Prior CABG, Previous cardiac surgeries, Prior Aortic Valve procedure, Prior Non-Aortic Valve procedure, Degenerative AV Etiology, Tricuspid Valve Morphology, Aortic Valve Insufficiency, Mitral Valve Insufficiency, Tricuspid Valve Insufficiency and Acuity.

2d1.2. What were the statistical results obtained from the analysis of the components? (e.g., *correlations, contribution of each component to the composite score, etc.*; ***if no empirical analysis****, identify the components that were considered and the pros and cons of each*)

All 4 non-fatal periprocedural complications in the composite were found to be associated with increased risk of one year mortality. Increased mortality risk was observed in patients with perioperative stroke (adjusted HR 2.10; 95% CI 1.65 to 2.87; p<0.001), major or life-threatening bleeding (adjusted HR 1.92; 95% CI 1.42 to 2.60, p<0.001), modified AKIN Stage III acute kidney injury (adjusted HR 1.81; 95% CI 1.38 to 2.37, p<0.001), and moderate or severe peri-valvular aortic regurgitation (adjusted HR 1.50; 95% CI 1.24 to 1.81; p<0.001).

Similarly, non-fatal periprocedural complications in the composite were also associated with 1-year patient reported health status as assessed by the KCCQ-OS score. Any stroke (adjusted impact on 1-year KCCQ-OS -5.8 points; 95% CI -9.2 to -2.4, p<0.001) and moderate or severe paravalvular regurgitation (adjusted KCCQ-OS impact -2.0 points; 95% CI-3.8 to -0.30, p=0.021) were independently associated with poorer adjusted KCCQ-OS at one year. Modified AKIN Stage III acute kidney injury (adjusted KCCQ-OS impact -3.3 points; 95% CI -6.8 to 0.28, p=0.07) and major or life-threatening bleed (adjusted KCCQ-OS impact 0.4 points; 95%CI -2.0 to 1.2, p=0.619) were not associated with one year KCCQ-OS but were retained in the global rank composite measure, given their strong associations with 1-year mortality.

2d1.3. What is your interpretation of the results in terms of demonstrating that the components included in the composite are consistent with the described quality construct and add value to the overall composite? (i.*e., what do the results mean in terms of supporting inclusion of the components;* ***if no empirical analysis****, provide rationale for the components that were selected)*

Empirical analyses confirm the importance of the selected endpoints by demonstrating their significant associations with one-year mortality and functional status.

2d2. Empirical analysis demonstrating that the aggregations and weighting rules are consistent with the quality construct and achieve the objective of simplicity to the extent possible

2d2.1 Describe the method used (*describe the steps—do not just name a method; what statistical analysis was used;* ***if no empirical analysis****, provide justification)*

Validation of endpoint rankings

The composite measure is based on a global ranking endpoint in which patients are classified according to the worst outcome (lowest rank score) that they experience. The rank ordering of endpoints in the composite was empirically validated by considering their associations with 1-year mortality and KCCQ.

Associations of component endpoints with categories based on overall composite

Note: This information is also presented in Section 2b1.2.

When endpoints in a composite occur with different frequencies, the composite may tend to be heavily influenced by the relatively more frequent endpoints, which can be problematic if the relatively more frequent endpoints have lesser clinical importance and if the less important endpoints dominate. In the case of the TVT

TAVR composite, there is not a major concern about an endpoint of lesser clinical importance dominating because all included endpoints have established associations with 1-year outcomes and the frequencies of the different endpoints are mostly of the same order of magnitude. However, it may still be concerning if sites identified as having high or low performance on the composite do not differ substantially with respect to one or more of the included endpoints. To assess this, we used the concept of performance categories. Hospitals were labeled as having better than expected performance if the 95% probability interval surrounding their composite score (site difference) fell entirely above 0, as performing worse than expected if the 95% probability interval surrounding their composite score (site difference) fell entirely below 0, and as expected otherwise. We compared risk-adjusted mortality and complication rates for endpoints in the composite across the three performance groups. We think it is desirable if the three groups have different risk-adjusted performance on each individual metric, as expected.

2d2.2. What were the statistical results obtained from the analysis of the aggregation and weighting rules? (e.g., *results of sensitivity analysis of effect of different aggregations and/or weighting rules; **if no empirical analysis**, identify the aggregation and weighting rules that were considered and the pros and cons of each*)

Validation of endpoint rankings

As shown in the table, endpoints in the composite were ranked in order of their decreasing hazard ratios with one-year mortality.

Endpoint Ranking	Component of Composite (Perioperative Complication)	Adjusted Hazard Ratio for One Year Mortality*
1	30-day death	N/A or ∞
2	30-day stroke	2.10
3	Major bleeding	1.92
4	Acute kidney injury	1.81
5	Moderate/severe peri-valvular leak	1.50

* Hazard ratio comparing patients who do versus do not experience perioperative complication

Associations of component endpoints with categories based on overall composite

Note: This information is also presented in Section 2b1.3.

Adjusted observed to expected (O/E) ratios of the individual endpoint components according to the 3 levels of site performance are summarized in the table. Sites with better than expected performance on the global rank composite metric showed lower O/E ratios for all components of the global rank composite measure compared with the sites that performed as expected or worse than expected. Similarly, sites with worse than expected performance on the global rank composite demonstrated consistently higher O/E ratios than the other sites. The largest differences favoring the better than expected sites were observed in the incidence of major, life threatening or disabling bleeding and moderate or severe paravalvular leak.

Variables	Observed to Expected Ratios Better than Expected (Sites = 25) (N = 7993)	Observed to Expected Ratios As Expected (Sites = 242) (N = 37473)	Observed to Expected Ratios Worse than Expected (Sites = 34) (N = 7095)
Death	0.71 (2.25% / 3.16%)	1.01 (3.21% / 3.17%)	1.25 (4.06% / 3.26%)
Stroke	0.73	1.03	1.29

Variables	Observed to Expected Ratios Better than Expected (Sites = 25) (N = 7993)	Observed to Expected Ratios As Expected (Sites = 242) (N = 37473)	Observed to Expected Ratios Worse than Expected (Sites = 34) (N = 7095)
	(1.74% / 2.38%)	(2.49% / 2.41%)	(3.16% / 2.44%)
Major or Life Threatening /Disabling Bleed	0.45 (2.84% / 6.30%)	1.02 (6.48% / 6.33%)	2.13 (13.6% / 6.38%)
Acute Kidney Injury or New Dialysis	0.67 (0.83% / 1.23%)	1.12 (1.34% / 1.20%)	1.17 (1.44% / 1.23%)
Moderate or Severe Peri-valvular Regurgitation	0.77 (1.88% / 2.45%)	1.19 (2.71% / 2.28%)	2.00 (4.76% / 2.38%)

2d2.3. What is your interpretation of the results in terms of demonstrating the aggregation and weighting rules are consistent with the described quality construct? (i.*e., what do the results mean in terms of supporting the selected rules for aggregation and weighting;* ***if no empirical analysis****, provide rationale for the selected rules for aggregation and weighting*)

Validation of endpoint rankings

Empirical analyses confirm that endpoints were ranked in order of their estimated empirical associations with one-year mortality.

Associations of component endpoints with categories based on overall composite

The test results show wide differences in risk-adjusted mortality and morbidity rates across categories of composite performance. These results support the validity of the composite measure as a quality measure for TAVR.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

If other:

3b. Electronic Sources

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

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

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

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

Attachment:

3c. Data Collection Strategy

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

3c.1. *Required for maintenance of endorsement*. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

***IF instrument-based,* consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.**

3c.2. Describe any fees, licensing, or other requirements to use any aspect of the measure as specified *(e.g., value/code set, risk model, programming code, algorithm)*.

1. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
*	*

*cell intentionally left blank

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

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

4a1.2. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing) what are the reasons? (*e.g., Do policies or actions of the developer/steward or accountable entities restrict access to performance results or impede implementation?*)
4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

4a2.1.1. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

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

4a2.1.2. Describe the process(es) involved, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

4a2.2.1. Summarize the feedback on measure performance and implementation from the measured entities and others described in 4d.1.

Describe how feedback was obtained.

4a2.2.2. Summarize the feedback obtained from those being measured.

4a2.2.3. Summarize the feedback obtained from other users

4a2.3. Describe how the feedback described in 4a2.2.1 has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

Improvement

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

4b1. Refer to data provided in 1b but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included.)

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

4b2. Unintended Consequences

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

4b2.1. Please explain any unexpected findings (positive or negative) during implementation of this measure including unintended impacts on patients.

4b2.2. Please explain any unexpected benefits from implementation of this measure.

2. Comparison to Related or Competing Measures

If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the

same target population), the measures are compared to address harmonization and/or selection of the best measure.

5. Relation to Other NQF-endorsed Measures

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

5.1a. List of related or competing measures (selected from NQF-endorsed measures)

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

5a. Harmonization of Related Measures

The measure specifications are harmonized with related measures;

OR

The differences in specifications are justified

5a.1. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s):

Are the measure specifications harmonized to the extent possible?

Yes

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

#2561: STS Aortic Valve Replacement (AVR) Composite Score

5b. Competing Measures

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

OR

Multiple measures are justified.

5b.1. If this measure conceptually addresses both the same measure focus and the same target population as NQF-endorsed measure(s):

Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality); OR provide a rationale for the additive value of endorsing an additional measure. (Provide analyses when possible.)

Appendix

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

Attachment:

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): America College of Cardiology

Co.2 Point of Contact: Jarrott, Mayfield, Jmayfield@acc.org

Co.3 Measure Developer if different from Measure Steward: American College of Cardiology

Co.4 Point of Contact: Susan, Fitzgerald, sfitzger@acc.org

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

Provide a list of sponsoring organizations and workgroup/panel members' names and organizations. Describe the members' role in measure development.

- Nimesh D Desai MD PhD – Hospital of the University of Pennsylvania (STS co-chair)
- David J Cohen MD MSc –University of Missouri-Kansas City (ACC co-chair)
- Sean M O'Brien PhD – Duke Clinical Research Institute (lead statistician)
- Sreekanth Vemulapalli MD -Duke Clinical Research Institute
- Suzanne V Arnold MD MHA - Saint Luke's Mid America Heart Institute, University of Missouri–Kansas City
- John K Forrest MD – Yale University
- Ajay J Kirtane MD, Columbia University
- Brian O'Neil MD, Henry Ford Medical Center
- Pratik Manandhar MS, Duke Clinical Research Institute
- David M Shahian MD – Massachusetts General
- Vinay Badhwar MD – University of West Virginia
- Vinod H Thourani MD – Piedmont Hospital
- John Carroll MD – University of Colorado
- Joseph E Bavaria MD - Hospital of the University of Pennsylvania

STS/ACC TVT Registry Risk Model Workgroup

(model developers)

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2020

Ad.3 Month and Year of most recent revision:

Ad.4 What is your frequency for review/update of this measure? No formal review scheduled at this time.

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

Ad.6 Copyright statement: American College of Cardiology Foundation All Rights Reserved

Ad.7 Disclaimers: STS and ACC do not have a web page dedicated to the TVT Registry measure specification. Participants can access a risk model companion guide to help them understand the model. The manuscript is also a publicly available resource.

Ad.8 Additional Information/Comments: STS and ACC appreciate the opportunity to submit measures for this NQF endorsement maintenance project.