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 #: 1550

Corresponding Measures:

De.2. Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: The measure estimates a hospital-level risk-standardized complication rate (RSCR) associated with elective primary THA and TKA in Medicare Fee-For-Service beneficiaries who are age 65 and older. The outcome (complication) is defined as any one of the specified complications occurring from the date of index admission to 90 days post date of the index admission (the admission included in the measure cohort).

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized complication rates (RSCRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA complications is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting complication rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. In addition, it has the potential to lower health care costs associated with complications.

S.4. Numerator Statement: The outcome for this measure is any complication occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. Complications are counted in the measure only if they occur during the index hospital admission or during a readmission. The complication outcome is a dichotomous (yes/no) outcome. If a patient experiences one or more of these

complications in the applicable time period, the complication outcome for that patient is counted in the measure as a "yes".

S.6. Denominator Statement: The target population for the publicly reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures.

Additional details are provided in S.7 Denominator Details.

S.8. Denominator Exclusions: This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare;
2. Who were discharged against medical advice (AMA); or,
3. Who had more than two THA/TKA procedure codes during the index hospitalization.

After applying these exclusion criteria, we randomly select one index admission for patients with multiple index admissions in a calendar year. We therefore exclude the other eligible index admissions in that year.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Enrollment Data

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 25, 2017

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is not formally paired with another measure; however, it is harmonized with the hospital-level, risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA), as well as the hospital-level, risk-standardized payment (RSP) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) measures.

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Summary of prior review in 2017

- This hospital-level, claims-based, outcome measure calculates a hospital-level risk-standardized complication rate (RSCR) associated with elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) in Medicare Fee-For-Service (FFS) beneficiaries who are age 65 and older.
- As a rationale for measuring this health outcome, the developer included a logic model that suggests that improved communication between providers involved at care transitions, prevention of and response to complications, patient safety, coordinated transitions to the outpatient environment, medication reconciliation, patient education, and disease management strategies leads to improved patient outcome by decreasing the risk of complications following elective primary THA and/or TKA.
- The developer submitted evidence for the measure noting the rates of complication and death following THA and TKA.

Changes to evidence from last review

☐ **The developer attests that there have been no changes in the evidence since the measure was last evaluated.**

☒ **The developer provided updated evidence for this measure:**

Updates:

- The developer provided updated citations for the rationale for measure development. The developer also included more recent studies that provide additional support for the previous conclusions.
    - The developer submitted updated evidence showing projected trend in performance of THA and TKA procedures and an increasing projected cost associated with performing both of these procedures. The developer provided data that shows variation in complication rates across hospitals, indicating there is room for quality improvement and targeted efforts to reduce these complications could result in better patient care and potential cost savings.

Exception to evidence

- N/A

Question for the Committee:

- *Is there at least one thing that the hospitals can do to achieve a change to improve patient health outcome following elective THA and/or TKA?*
- *The evidence provided by the developer is updated, directionally the same, and stronger compared to that for the previous NQF review. Does the Committee agree there is no need for repeat discussion and vote on Evidence?*

Guidance from the Evidence Algorithm

The measure assesses performance on a health outcome of decreasing complication rates (box 1) → The relationship between decreased risk of complications and a hospital's quality of care is demonstrated through empirical data (box 2) → Pass

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

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

Maintenance measures – increased emphasis on gap and variation

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

- The developers provided three-year, hospital-level, RSCR from April 1, 2016 to March 31, 2019 using Medicare administrative claims data (n= 962,744 admissions) from 3,418 hospitals.
    - The RSCRs have a mean of 2.5% and range from 1.2-10.6% in the study cohort. The median risk-standardized rate is 2.4%.
- The developers also presented the distribution of RSCRs across hospitals over a three-year period.

Disparities

- The developer provided disparities data on THA/TKA RSMR across hospitals by proportion of patients with social risk (dual-eligible patients and AHRQ SES Index Scores).
    - RSCR across hospitals by proportion of dual eligible patients: Data was provided from the national Medicare FFS claims and Medicare Beneficiary Summary File (MBSF) for a period of performance from April 1, 2016 to March 31, 2019
    - RSCR across hospitals by Proportion of Patients with AHRQ SES Index Scores: Data was provided from the national Medicare FFS claims and the American Community Survey (ACS); Medicare FFS claims data was for a period of performance from April 1, 2016 to March 31, 2019 for hospitals; and ACS data was for a period of performance from 2013-2017.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Are you aware of evidence that additional disparities exist in this area of healthcare aside from what the developer provided?

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

Committee Pre-evaluation Comments:
Criteria 1: Importance to Measure and Report (including 1a, 1b, 1c)

1a. Evidence to Support Measure Focus: 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.

evidence applies directly

This measure is still solid and updated evidence show room for improvement and variation among hospitals and patient populations.

Clinical guidelines and evidence continues to support.

Evidence is acceptable. The logic model is good.

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?

no concern

There continues to be performance gap and socioeconomic factors are important to tease out going forward.

Data shows variation in rates. Disparities due to social risk.

The developers reported that the median hospital 30-day, all-cause, RSCR for the THA/TKA complications measure for the 3-year period between April 1, 2016 – March 31, 2019 was 2.4%. The RSCRs have a mean of 2.5% and range from 1.2-10.6% in the study cohort. The median risk-standardized rate is 2.4%. The 10th and 90th percentile (high is bad) are 1.9 and 3.0%. From the SMP: "There is some variation in the calculated scores; however, the statistical choice of how to categorize hospitals into 3 performance categories leaves 96% of hospitals in "no different from the U.S. national rate", which does not reflect much variation". Can the developers describe a level of performance and variation in performance that would justify retiring the measure?

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

Reliability

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

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

Validity

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

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

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 (SMP) Subgroup

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the NQF Scientific Methods Panel (SMP). The Subgroup passed the measure on reliability and validity. The measure was not pulled for discussion during the October 2020 meeting. A summary of the measure and the Panel's review is provided below.

Reliability

- The SMP passed the measure on reliability with Moderate rating (H-2; M-6; L-0; I-0).
- The developers conducted two types of reliability testing. The developers estimated the measure score level by calculating the intra-class correlation coefficient (ICC) using a split sample (test-retest) method, and then estimated the facility-level reliability (signal-to-noise reliability) using Adams' Method.
    - For signal-to-noise analysis, the developers reported a median reliability of 0.87, ranging from 0.46 to 1.00, and a mean of 0.83. The 25th and 75th percentiles were 0.74 and 0.94, respectively. The developers noted that the median reliability score demonstrates moderate reliability.
    - For split-sample reliability, the developers included 962,744 admissions in the analysis using three years of data. As a metric of agreement, the developers calculated the ICC for hospitals with 25 admissions or more. Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSCR for each hospital was 0.524. The developer noted that this is a lower bound.
- The SMP reviewers generally agreed that the testing approach and results were acceptable.

Validity

- The SMP passed the measure on validity with Moderate rating (H-0; M-6; L-1; I-1).
- The developers conducted validity testing at the measure score level. The measure was compared to the Overall Hospital Star Rating and Hospital THA/TKA Surgical Volume.
    - The developers reported the correlation between THA/TKA complications and Star-Rating summary score to be -0.185, which suggests that hospitals with lower THA/TKA RSCRs are more likely to have higher Star-Rating summary scores especially at the extremes.
    - A general trend was noted that high-volume hospitals (i.e., those in the upper deciles) have lower RSCRs than hospitals in other volume deciles.
    - Developer stated that overall, the results above show that the trend and direction of this association is in line with what would be expected. Risk model discrimination and calibration: c statistic = 0.65; Developer reports good discrimination and predictive ability based on risk decile plot.
- The SMP reviewers generally accepted the validity testing results as a weak but acceptable demonstration of validity.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

Committee Pre-evaluation Comments:
Criteria 2: Scientific Acceptability of Measure Properties (including all 2a, 2b, and 2c)

2a1. Reliability-Specifications: 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?

no concerns

The measure can be consistently implemented. Appropriate description and exclusions.

No concerns with reliability.

Specifications are fine

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

no concerns

None. Scientific Methods Panel rated reliability moderate.

No concerns with reliability.

I accept the judgements of the SMP. Reliability is relatively strong in methods and results. Because the scores are transformed into star ratings, I would have like to see some analysis on the stability (reliability) of those categories.

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

no concern

None. Scientific Methods Panel satisfied with validity.

No concerns with validity.

Validity testing correlating score with star rating summary score is weak. Correlating a measure to a rating scale of which it is a component is not very compelling. The logic model is good and could have provided the basis for better validity analyses. Measure does not distinguish been one or many complications per patient. So entities with more complications per patient (who have any) will not be distinguishable from entities with fewer complications per patient (who have any).

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?

all patients are medicare fee for service

Exclusions and risk adjustment are appropriate and justified.

No issues.

No problems

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?

could high volume hospitals affect the validity results

None

No major concerns.

What is the distribution of performance in the 1 star entities? I'm curious how different the best performance in this group differs from the national average.

Criterion 3. Feasibility

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

3. **Feasibility** is the extent to which the specifications including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.
   - All data elements for this measure originate from defined fields in electronic claims.
   - The data is generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score). It is coded by someone other than the person obtaining original information (e.g., DRG, ICD-9 codes on claims)
   - The developer noted that this measure uses administrative claims and enrollment data and as such, it offers no data collection burden to hospitals or providers

Questions for the Committee:

- 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?

data elements are defined in electronic claims

Electronic data coded from patient care data. Easily accessible.

No issues or concerns with feasibility.

The measure has been in use for years and appears feasible.

Criterion 4: Usability and Use

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

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

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

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

Current uses of the measure

Publicly reported? ☒ **Yes** ☐ **No**

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

Accountability program details

- The developer noted that the measure is publicly reported on CMS' Care Compare website. Under Care Compare and other CMS public reporting websites, CMS collects quality data from hospitals, with the goal of driving quality improvement through measurement and transparency by publicly displaying data to help consumers make more informed decisions about their health care. It is also intended to encourage hospitals and clinicians to improve the quality and cost of inpatient care provided to all patients. The data collected are available to consumers and providers on the Care Compare website
- The developer noted that the measure is also used for paying a portion of hospitals based on the quality and efficiency of care as part of CMS' Hospital Value-Based Purchasing (HVBP) Program.
- The developer added that the exact number of measured entities (acute care hospitals) varies with each new measurement period. For the period between 2016 – 2019, all non-federal short-term acute care hospitals (including Indian Health Service hospitals) and critical access hospitals (3,418 hospitals) were included in the measure calculation. Only those hospitals with at least 25 THA/TKA procedures were included in public reporting.

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. Those being measured receive performance results and data via CMS' QualityNet website. The website also contains detailed patient-level results and benchmarks to assist in interpretation.
2. The developer noted that measured entities can submit feedback about the measure through an email inbox. Experts on measure specifications, calculation, or implementation, prepare responses to those inquiries and reply directly to the sender.
3. The developers state that they consider feedback when reevaluating measures. The developers state that they have not received any feedback from stakeholders that would require additional analysis or changes to the measure since the last endorsement maintenance cycle.

Additional Feedback:

- The developer noted that since the last endorsement cycle, they have reviewed more than 500 articles related to 90-day complications following an elective THA/TKA procedure. Relevant articles shared key

themes related to additional risk variables for consideration, including social risk factors and other clinical comorbidities; outcome rate and risk variable comparisons between inpatient and outpatient settings for both TKA and THA procedures; exploration of potential association between length of stay and THA/TKA complication outcome rates; and the relationship between complication rates and costs of care.

***Questions for the Committee**:*

- 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 developers reported that the median hospital 30-day, all-cause, RSCR for the THA/TKA complications measure for the 3-year period between April 1, 2016 – March 31, 2019 was 2.4%.
- The median RSCR decreased by 0.1 absolute percentage points from April 2016 – March 2017 (median RSCR: 2.5%) to April 2018 – March 2019 (median: RSCR: 2.4%).
- The developer included recent peer-reviewed literature (Bozic et al., 2020) to confirm the trends reported above.

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

Unexpected findings (positive or negative) during implementation

- The developers reported that they have not seen any unexpected findings.

Potential harms

- The developer noted that potential harms could include providers inappropriately shifting care in response to this measure, increased patient morbidity and mortality, and other unintended consequences for patients. The developers monitor for these unintended consequences and have not seen any indications that it is occurring.

Additional Feedback:

- N/A

***Questions for the Committee**:*

- How can the performance results be used to further the goal of high-quality, efficient healthcare?
- Do the benefits of the measure outweigh any potential unintended consequences?

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

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

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?

yes

Used in Hospital Compare and CMS websites. Used for Value Based Purchasing Program. Many recent publications using data collected in order to look at risk mitigation in these patients.

No concerns with use; publicly reported as part of accountability program. Feedback utilized.

The measure is in use

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.

benefits outweigh harm potential

No negative findings. Benefits shown as gradual improvement over time in measure.

No usability issues.

The median RSCR decreased by 0.1 absolute percentage points from April 2016-March 2017 (median RSCR: 2.5%) to April 2018-March 2019 (median: RSCR: 2.4%). But the performance of the higher percentiles has been stable. It is not clear that performance results have driven improvements. Is there evidence that entities with below average performance have actively made changes related to subsequent improvements?

Criterion 5: Related and Competing Measures

Related or competing measures

- The developer identified two related measures in their original submission:
  - NQF #1551 Hospital-Level 30-Day Risk-Standardized Readmission Rate (RSRR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA)
  - NQF #3493 Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) for Merit-Based Incentive Payment System (MIPS) Eligible Clinicians and Eligible Clinician Groups
- The developer later identified an additional related measure:
  - NQF #3474 Hospital-Level, Risk-Standardized Payment Associated With a 90-Day Episode of Care for Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

Harmonization

- The developer stated that the measure specifications are harmonized to the extent possible. They noted that they focused on related outcome (mortality and readmissions) measures in their harmonization analysis. Their rationale for this was that clinical coherence of the measured cohort takes precedence over alignment with related non-outcome measures. They state that many process measures are limited due to the broader patient exclusions necessary to examine only a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

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?

no competing measures, related measure with 3493

Yes. Related measures properly addressed and harmonized.

Few related measures, no concerns with harmonization.

no comments

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: 01/26/2021

Comment by: American Medical Association

The American Medical Association (AMA) appreciates the opportunity to comment on the NQF Measure #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). In reviewing the calculation, we are disappointed to see the minimum measure score reliability result calculated at 0.46 and the intraclass correlation coefficient (ICC) calculated at 0.524 using a minimum case number of just 25 patients. We believe that measures must meet minimum acceptable thresholds of 0.7 for reliability and require higher case minimums to allow the overwhelming majority of hospitals to achieve an ICC of 0.6 or higher.

The AMA is also extremely concerned that the measure developer used the recommendation to exclude social risk factors in the risk adjustment models for measures that are publicly reported as outlined in the recent report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program (ASPE, 2020). We believe that while the current testing may not have produced results that would indicate incorporation of the two social risk factors included in testing, this measure is currently used both for public reporting and value-based purchasing. A primary limitation of the ASPE report was that none of the recommendations adequately addressed whether it was or appropriate to adjust for social risk factors in the same measure used for more than one accountability purpose, which is the case here. This discrepancy along with the fact that the additional analysis using the American Community Survey is not yet released must be addressed prior to any reliance on the recommendations within this report.

In addition, we question whether the measure continues to be useful to distinguish hospital performance and drive improvements based on the distribution of hospital's performance scores where only 60 hospitals performed better than the national rate and 50 hospitals performed worse (as noted in section 2b4 and the discussion on improvement in section 4b1 of the measure submission form), and where there was only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019.

We request that the Standing Committee evaluate whether the measure continues to meet the measure evaluation criteria required for endorsement.

Reference:

Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs

Comment by: Federation of American Hospitals

The Federation of American Hospitals (FAH) appreciates the opportunity to comment on Measure #1550, Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). The FAH is concerned that even though the median reliability score was 0.87 for hospitals with at least 25 cases, reliability ranged from 0.46 to 1.00 and that the intraclass correlation coefficients (ICC) was 0.524. The FAH believes that the developer must increase the minimum sample size to a higher number to produce a minimum reliability threshold of sufficient magnitude (e.g. 0.7 or higher) and an ICC of 0.6 or higher.

In addition, the FAH is very concerned to see that the measure developer's rationale to not include social risk factors in the risk adjustment model was in part based on the recommendations from the report to Congress by Assistant Secretary for Planning and Evaluation (ASPE) on Social Risk Factors and Performance in Medicare's Value-based Purchasing program released in March of last year (ASPE, 2020). A fundament flaw within the ASPE report was the lack of any recommendation addressing how a single measure with multiple accountability uses should address inclusion of social risk factors as is the case with this measure, which is both publicly reported and included in the Hospital Value-Based Purchasing program. Regardless of whether the testing of social risk factors produced results that were sufficiently significant, the FAH believes that no developer should rely on the recommendations of this report until the question of how to handle multiple uses is addressed along with the additional analysis using the American Community Survey.

Lastly, the FAH is concerned that there is insufficient variation in performance across hospitals and limited opportunities for improvement to support this measure's continued use in accountability programs. Specifically, the performance scores reported in 2b4. Identification of Statistically Significant and Meaningful Difference in Performance are generally low with only 60 hospitals identified as better than the national rate and 50 are worse than the national rate. We base our concerns on these results along with the discussion on improvement in section 4b1 of the measure submission form where only an increase of 0.1 absolute percentage points between July 2016-June 2017 and July 2018-June 2019 was found.

As a result, the FAH requests that the Standing Committee carefully consider whether the measure as specified should continue to be endorsed.

Reference:

Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health & Human Services. Second Report to Congress on Social Risk Factors and Performance in Medicare's Value-Based Purchasing Program. 2020. https://aspe.hhs.gov/social-risk-factors-and-medicares-value-based-purchasing-programs

Of the 1 NQF member who has submitted a support/non-support choice:

0 support the measure

1 does not support the measure

Combined Methods Panel Scientific Acceptability Evaluation

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 1550

Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

Type of measure:

☐ **Process** ☐ **Process: Appropriate Use** ☐ **Structure** ☐ **Efficiency** ☐ **Cost/Resource Use**

☒ **Outcome** ☐ **Outcome: PRO-PM** ☐ **Outcome: Intermediate Clinical Outcome** ☐ **Composite**

Data Source:

☒ Claims ☐ Electronic Health Data ☐ Electronic Health Records ☐ Management Data
☐ Assessment Data ☐ Paper Medical Records ☐ Instrument-Based Data ☐ Registry Data
☒ Enrollment Data ☒ Other

Level of Analysis:

☐ Clinician: Group/Practice ☐ Clinician: Individual ☒ Facility ☐ Health Plan
☐ Population: Community, County or City ☐ Population: Regional and State
☐ Integrated Delivery System ☐ Other

Measure is:

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

RELIABILITY: SPECIFICATIONS

1. **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.*

2. **Briefly summarize any concerns about the measure specifications.**

   Panel Member 1: No concerns

   P**anel Member 2:** None

   Panel Member 3: None

   Panel Member 4: None

   Panel Member 5 : I find it confusing that the measure is specified in the brief description and in the denominator as patients 65+, but then also for patients 18+. The denominator specified only 65+. Please clarify.

   As mentioned above, the specifications include patients aged 18+ with a statement that the measure has been tested in both patients aged 18 years and older and those aged 65 years or older. However, it seems that most testing was conducted using data from patients aged 65+. The only testing conducted for patients aged 18-64 vs. 65+ was for the risk-adjustment model (section 2b3.11) using data from 2006. I could not identify any other testing of reliability, validity, threats to validity, or performance that included data for the younger age group. This questions the reliability, and possibly also the validity of this measure for patients aged 18-64. This issue has been clarified, and measure developers decided to change the specifications to limit each of the measures to the Medicare FFS 65+ population. The ratings for reliability and validity were selected accordingly.

   Panel Member 6: None

   Panel Member 7: None

   Panel Member 8: None

RELIABILITY: TESTING

Submission document: "MIF_1550" document for specifications, testing attachment questions 1.1-1.4 and section 2a2

3. **Reliability testing level ☒ Measure score ☐ Data element ☐ Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**
   ☒ Yes ☐ No
5. 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
6. **Assess the method(s) used for reliability testing**

   Submission document: Testing attachment, section 2a2.2

   Panel Member 1: Developer used a split sample ICC and signal to noise approaches, which were appropriate.

   Panel Member 2: Split sample ICC – test/retest. Signal/noise analysis at the hospital level using HLR

   Panel Member 3: Used standard methods:

   SNR (0.87) and split-sample reliability testing (0.52) consistent with acceptable reliability.

   Panel Member 4: Used two appropriate methods for testing – split sample and signal-to-noise.

   Panel Member 5: No concerns. Methods were appropriate and clearly described.

   A description of how the 25-case threshold for public reporting was determined would be useful.

   Panel Member 6: Split sample and signal-to-noise; appropriate

   Also, chart review (although developers included this in the validation data)

   Panel Member 7: ICC using a split sample (test-retest) method then estimated the facility-level reliability (signal-to-noise reliability).

   Panel Member 8: The developers performed two types of reliability testing. First, they estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (i.e. test-retest) method. Second, they estimated the facility-level reliability (signal-to-noise reliability).
7. **Assess the results of reliability testing**

   Submission document: Testing attachment, section 2a2.3

   Panel Member 1: The STN analysis found a median facility reliability estimate of 0.87 which was quite strong, however the split sample ICC demonstrated a facility-level reliability estimate of 0.52, which is suboptimal at best. Therefore, we have a mixed picture of reliability. Conceptually, I find the STN more compelling as a measure of reliability so I am inclined to weight those results more heavily than the ICC method using the test-retest framework.

   Panel Member 2: Split sample correlation is 0.524 – low considering the sample size of 3,365 hospitals. The submitter claims moderate; I think that is an overstatement. Signal to noise shows moderate reliability also.

   Panel Member 3: SNR (0.87) and split-sample reliability testing (0.52) consistent with acceptable reliability.

   Panel Member 4: Median signal-to-noise score of 0.87, which demonstrates “almost perfect” agreement, as defined by Adams et al.

   Split-sample score was 0.524 is represents the lower bound of the true reliability.

   Panel Member 5: I don’t think the interpretation of SNR reliability estimate as an agreement statistic is appropriate. Results suggest acceptable reliability at the score level (>0.7), thus there is high/acceptable certainty that the performance measure scores are reliable. However, it is now known how the inclusion of patients below the age of 65 would have impacted these results.

It would be useful to report here the percent of hospitals included in the reliability results (25+ cases), although this is reported in the performance section (655/3418=19%).

Panel Member 6: Split sample ICC=0.524; Signal-to-noise mean 0.83 (0.46-1.00); reasonable evidence of reliability.

Additional historical chart review was performed to "validate" the method of using claims data to document complications; after adjustment of the complication definitions, 99% agreement was achieved

Panel Member 7: Split sample - 0.524. Median signal-to-noise - 0.87. What is underlying the difference? What is the implication for this? Which matters more?

Panel Member 8: Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSCR for each hospital was 0.524. The median SNR reliability score was 0.87, ranging from 0.46 to 1.00. The 25th and 75th percentiles were 0.74 and 0.94, respectively. It's curious that the split-sample and median SNR reliabilities are so different. Which is more important for evaluating the measure? How are they conceptually different? Although the Landis modifiers are cited, I do not accept them as relevant to this context. The Landis modifiers pertain to the strength of evidence against the null hypothesis of no agreement between raters of a categorical classifier. Thus I would rate the overall reliability as poor but the median facility reliability as very good. Note that other modifiers exist: Koo 2016 - "values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively. Portney and Watkins are more conservative, particularly at the upper end, with <0.75 poor to moderate, >0.75 good, an >0.90 ''reasonable for clinical measurements''.

I think we really need to move beyond these modifiers and do some work on the implications of unreliability in different quality measurement contexts. Can the developers comment of the impact of the observed reliability on misclassification or other consequences?

8. 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)

9. 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)

10. **OVERALL RATING OF RELIABILITY** (taking into account precision of specifications and <u>all</u> 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)

11. **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: See comments for item #7.

Panel Member 2: See #7

Panel Member 3: SNR (0.87) and split-sample reliability testing (0.52) consistent with acceptable reliability.

Panel Member 4: Used two appropriate methods for testing; signal-to-noise produced a score that demonstrated 'almost perfect' agreement.

Panel Member 5: Results suggest acceptable reliability at the score level, thus there is high certainty that the performance measure scores are reliable.

Can developers elaborate on how the 25-case threshold was established in relation to the overall reliability results?

Panel Member 6: Appropriate testing with moderate to high agreement

Panel Member 7: VALIDITY: ASSESSMENT SNR is OK enough…

Panel Member 8: See comments under #7

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

12. **Please describe any concerns you have with measure exclusions.**

Submission document: Testing attachment, section 2b2.

Panel Member 1: None

Panel Member 2: None

Panel Member 3: none

Panel Member 4: None

Panel Member 5 : No concerns. Exclusion rates were generally very low.

Panel Member 6: No concerns

Panel Member 7: None

Panel Member 8: None

13. **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: None

Panel Member 3: None

Panel Member 4 : There is variation in the calculated scores; however, the statistical choice of how to categorize hospitals into 3 performance categories leaves 96% of hospitals in "no different from the U.S. national rate", which does not reflect much variation.

Panel Member 5 : As noted above, a clarification about the patient level performance transformation would be helpful: "The results are then transformed and…".

As reported, 19% (655/3418) of hospitals had fewer than 25 cases therefor could not be reliably assessed for their RSMR (risk-standardized mortality rate). Can developers elaborate on how the 25-case threshold was established?

A concern arises from the fact that 96% [2,653/(2,653+60+50)] of hospitals assessed performed no different from the U.S. national rate. Even though a higher risk of readmission was noted for high risk hospitals compared to low risk hospitals, this was true for only 50/(2,653+60+50) of hospitals, i.e., <2%, which questions the usefulness of this measure which could be very close to being topped out.

Panel Member 6: Developers consider three levels of complications depending on the time frame but then simply consider a binary outcome of yes/no any complication. Therefore, a patient with an MI, pulmonary embolism, stroke and prosthetic infection would be given the same weight as another patient with pneumonia. There is certainly evidence from surgical literature that multiple complications are associated with higher mortality than single complications and various complications have different risk of mortality associated with them. To the extent that complications are associated with processes of care, the inability to distinguish between sites with multiple complications vs. those with primarily isolated complications is a major limitation.

Panel Member 7: None

Panel Member 8: None

14. **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: N/A

Panel Member 2: None

Panel Member 3: None

Panel Member 4: Not applicable

Panel Member 5: None

Panel Member 6: Multiple data sources (claims, eligibility, etc.) however ability to merge data is well established.

Panel Member 7: None

Panel Member 8: None

15. **Please describe any concerns you have regarding missing data.**

Submission document: Testing attachment, section 2b6.

Panel Member 1: None

Panel Member 2: None

Panel Member 3: none

Panel Member 4: None

Panel Member 5 : No concerns – no missing data reported

Panel Member 6: No major concern

Panel Member 7: None

Panel Member 8: None

16. **Risk Adjustment**

16a. **Risk-adjustment method** ☐ **None** ☒ **Statistical model** ☐ **Stratification**

16b. **If not risk-adjusted, is this supported by either a conceptual rationale or empirical analyses?**

☐ Yes ☐ No ☒ Not applicable

16c. **Social risk adjustment:**

16c.1 Are social risk factors included in risk model? ☒ Yes ☒ No ☐ Not applicable

16c.2 Conceptual rationale for social risk factors included? ☒ Yes ☐ No

16c.3 Is there a conceptual relationship between potential social risk factor variables and the measure focus? ☒ Yes ☐ No

16d. **Risk adjustment summary:**

16d.1 All of the risk-adjustment variables present at the start of care? ☒ Yes ☐ No

16d.2 If factors not present at the start of care, do you agree with the rationale provided for inclusion?
☐ Yes ☐ No

16d.3 Is the risk adjustment approach appropriately developed and assessed? ☒ Yes ☐ No

16d.4 Do analyses indicate acceptable results (e.g., acceptable discrimination and calibration)
☒ Yes ☐ No

16d.5. Appropriate risk-adjustment strategy included in the measure? ☒ Yes ☒ No

16e**. Assess the risk-adjustment approach**

Panel Member 1: The risk adjustment approach is sound and results are acceptable.

Panel Member 2: Typical of claims based measures – using HCCs and demographics. C-statistic is consistent with previous submission (0.65)

Panel Member 3: Hierarchical logistic regression model with PE ratio.

Discrimination is acceptable. (C stat 0.65) and calibration is acceptable (0.04, 1.02)

Panel Member 4 : Used hierarchical logistic regression model; c-statistic of 0.65, which indicates moderate model discrimination

Panel Member 5 : I have a few concerns, and would appreciate if developers could address the following issues:

1. Interpretation of Table 5 (Adjusted OR and 95% CIs for the AMI Mortality Hierarchical Logistic Regression Model over Different Time Periods in the Testing Dataset), notable for Osteoarthritis that was associated with lower risk of complications. Could this be due to collinearity with other risk-factors?

2. The conclusions drawn from the results of the estimation of average hospital and patient effects related to social risk factors (decomposition analyses) was not clear to me. If I am interpreting correctly, and as also noted by the developers, the patient-level effect was greater than the hospital-level effect for dual eligible status. Thus, dual eligibility acts in a similar manner compared to other clinical factors assesses. This was different than the results presented for the AHRQ SES index. Following the logic presented, shouldn't this support including Dual-eligibility in the risk-adjustment model?

3. The decision to not include social risk factors in the model is supported mainly by testing results of no added predictive power and no change in hospital performance rankings. It would be useful to know the rate of hospitals that would have change rank if social-risk factors would have been included, which would provide information on the practical implication not informed by a correlation coefficient between RSRRs for each hospital with and without dual eligibility added. Regarding the result of no added predictive power, have similar considerations been applied to significant clinical factors included in the model, or even more, to non-significant clinical factors which are also expected to have no impact on the model's predictive power and hospital ranking?

Panel Member 6:

1) Risk adjustment model hs c-statistic of 0.65, which is weak, especially for a publicly reportable measure.
2) Binary outcome which does not distinguish between single and multiple complications (leave aside the thorny problem of "weighting" individual complications) may represent institutional quality (or lack thereof). Developers would need to prove that the distribution of complications is equivalent across institutions—i.e., among institutions who had a complication, equal proportions had one, two, three or more complications—that there were not some institutions that had mor multiple complications than others

Issue of social risk factors problematic—after presenting lengthy rationale for including them, as well as documentation of their impact on outcome authors essentially make a political rather than methodological decision not to include them. That the hospital level effects were significant indicates that

if the dual eligible or low AHRQ SES Index variables were used in the model to adjust for patient-level differences, then some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality….ASPE's latest report to Congress highlights which SRFs are valid in claims data, and that adjustment for SRFs in publicly reported quality measures is not recommended because providers should be accountable for overall outcomes, regardless of social risk (ASPE 2020). In other words, social risk factors do impact outcome but developers are choosing not to account for them in the metric of hospital quality If a hospital cares for a frail patient, outcomes are adjusted; but if a hospital cares for a patient whose proclivity to complication is increased because of inadequate primary care or social support, the same hospital's outcome is not adjusted. The rationale baffles comprehension.

Panel Member 7: Methods are fine, as are discrimination and calibration.

Panel Member 8: Adequate method, discrimination, and calibration.

For cost/resource use measures ONLY:

17. **Are the specifications in alignment with the stated measure intent?**

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

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

VALIDITY: TESTING

19. **Validity testing level:** ☒ **Measure score** ☐ **Data element** ☐ **Both**
20. **Method of establishing validity of the measure score:**

    ☒ **Face validity**

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

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

21. **Assess the method(s) for establishing validity**

    Panel Member 1: The developer used a reasonable approach of correlating the measure score with CMS's overall Hospital Star Rating and a measure of joint replacement surgical volume.

    Panel Member 2: Correlations with the CMS's Overall Hospital Star Rating and TJA volume

    Panel Member 3: Assessed correlation with existing measures:

    Star rating summary score: correlation coefficient -0.185

    Hospital volume-outcome association: higher volume centers tended to have lower readmission rates

    Panel Member 4: For empirical validity testing, compared the hospital's performance on the TKA/THA complication measure to the hospital's overall Summary star rating and hospital volume. **Concerns with demonstrating validity by using a comparator measure that includes the measure being tested.** (we would expect there to be some correlation!)

    Panel Member 5 : Face validity was supported during the measure development phase based on national guidelines for publicly reported outcomes measures, and volume. the inclusion of consultation with outside experts and with the public.

    Empirical testing against other similar measures were appropriate.

    Panel Member 6: The data on the face validity of the expert panel is lacking. Specifically it would be important to know how they arrived at the binary outcome for multiple potential complications and what supported that decision.

    Correlation with CMS star ratings is questionable depending upon whether or not this metric is at all included in that system. Similarly, correlation with hospital volume would be meaningful if there were

data presented confirming that increased hospital volume correlates with fewer complications. Perhaps a more meaningful correlation might be made with mortality—a universally accepted metric of quality.

Since it is fairly universally accepted that complications relate to quality, the only issue which really needs validation here is the decision to use a single binary outcome rather than accounting for multiple complications.

Panel Member 7: They report correlations between CMS's Overall Hospital Star Rating and TJA complications and volume.

Panel Member 8: Correlations with the CMS's Overall Hospital Star Rating and TJA volume

22. **Assess the results(s) for establishing validity**

    Submission document: Testing attachment, section 2b2.3

    Panel Member 1: The measure score correlation with the overall star-rating was weak and in expected direction (-0.185), while the correlation with joint replacement volume was stronger and still in the expected direction (-0.26). The weak correlation with overall quality is troubling, but overall, these findings provide support for the validity of the measures

    Panel Member 2: The correlation with Star-Rating summary score is -0.185. Correlation with TJA volume was-0.256. Both are low.

    Panel Member 3: Approach/results are acceptable for establishing empiric validity

    Panel Member 4 : Low correlation (-0.185) with Summary star rating.

    Start to see some trend in higher volumes resulting in lower complication rate, but only at the highest volumes (70-100% deciles).

    Panel Member 5 : Empirical testing results suggest low level evidence of validity against other related measures (correlations under 0.3).

    Panel Member 6: None

    Panel Member 7: Higher stars, lower complications. Higher volumes, lower complications.

    Panel Member 8: The correlation between THA/TKA complications and Star-Rating summary score is -0.185. Correlation with TJA volume was-0.256.

23. **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)

24. **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)

25. **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.)

26. **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: The validity testing results were acceptable, although the weak correlation with overall quality is concerning.

Panel Member 2: The correlation with related measures is low – the validity testing in general is light in this application.

Panel Member 3: Assessed correlation with existing measures:

Star rating summary score: correlation coefficient -0.185

Hospital volume-outcome association: higher volume centers tended to have lower readmission rates

Panel Member 4 : Concerns with the choice of one of the measures chosen (Summary star rating) to empirically test this measure's validity; but, the addition of looking at the relationship between complication rate and volume was a good analytical choice and did demonstrate the expected relationship.

I do have concerns with the lack of variation in performance as displayed through the publicly reported performance categories (96% are 'no different than average').

Panel Member 5 : Results suggest low correlation with similar measures at the score level (<0.3), thus there is a low to moderate certainty that the performance measure scores are valid, with trends of associations shown to be in the expected directions. Face validity was supported.

Panel Member 6: Validity of use of single binary outcome regardless of how many complications and absence of inclusion of social risk factors require better validation. Testing for the >18 population done using 2006 data—needs to be updated.

Panel Member 7: I wish we had better ways to validate than CMS stars and the volume-outcome relationship. I don't have an answer – therefore moderate.

Panel Member 8: What magnitude of correlation would support or fail to support the validity hypothesis. With this sample size, statistical significance is not helpful. What does the very modest correlation with an overall quality index mean in terms of the quality of TJA treatment? Could the signal here just be picking up incomplete risk adjustment? What are the in-hospital processes that are hypothesized to be associated with TJA complications? Testing those hypotheses would be more convincing.

FOR COMPOSITE MEASURES ONLY: Empirical analyses to support composite construction

27. **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**

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

ADDITIONAL RECOMMENDATIONS

29. **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 #: 1550

Corresponding Measures:

De.2. Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

Co.1.1. Measure Steward: Centers for Medicare & Medicaid Services

De.3. Brief Description of Measure: The measure estimates a hospital-level risk-standardized complication rate (RSCR) associated with elective primary THA and TKA in Medicare Fee-For-Service beneficiaries who are age 65 and older. The outcome (complication) is defined as any one of the specified complications occurring from the date of index admission to 90 days post date of the index admission (the admission included in the measure cohort).

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized complication rates (RSCRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA complications is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting complication rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. In addition, it has the potential to lower health care costs associated with complications.

S.4. Numerator Statement: The outcome for this measure is any complication occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. Complications are counted in the measure only if they occur during the index hospital admission or during a readmission. The complication outcome is a dichotomous (yes/no) outcome. If a patient experiences one or more of these complications in the applicable time period, the complication outcome for that patient is counted in the measure as a "yes".

S.6. Denominator Statement: The target population for the publicly reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures.

Additional details are provided in S.7 Denominator Details.

S.8. Denominator Exclusions: This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare;
2. Who were discharged against medical advice (AMA); or,
3. Who had more than two THA/TKA procedure codes during the index hospitalization.

After applying these exclusion criteria, we randomly select one index admission for patients with multiple index admissions in a calendar year. We therefore exclude the other eligible index admissions in that year.

De.1. Measure Type: Outcome

S.17. Data Source: Claims, Enrollment Data

S.20. Level of Analysis: Facility

IF Endorsement Maintenance – Original Endorsement Date: Jan 31, 2012 **Most Recent Endorsement Date:** Jan 25, 2017

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is not formally paired with another measure; however, it is harmonized with the hospital-level, risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA), as well as the hospital-level, risk-standardized payment (RSP) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) measures.

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

NQF_evidence_THATKAcomplications_Fall2020_final_7.22.20.docx

1a.1 *For Maintenance of Endorsement*: Is there new evidence about the measure since the last update/submission?

Do not remove any existing information. If there have been any changes to evidence, the Committee will consider the new evidence. Please use the most current version of the evidence attachment (v7.1). Please use red font to indicate updated evidence.

Yes

1a. Evidence (subcriterion 1a)

NATIONAL QUALITY FORUM—Evidence (subcriterion 1a)

Measure Number (*if previously endorsed*)**:** 1550

Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

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

Date of Submission: **11/2/2020**

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

Outcome

☒ Outcome: **Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)**

☐ Patient-reported outcome (PRO):

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

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

☐ Process:

☐ Appropriate use measure:

☐ Structure:

☐ Composite:

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

Figure 1. THA/TKA Complications Logic Model

The goal of this measure is to improve patient outcomes by providing patients, physicians, and hospitals with information about hospital-level, risk-standardized complication rates following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA). Measurement of patient outcomes allows for a broader view of a hospital's quality of care that encompasses more than what can be captured by individual process of care measures. More specifically, complex and critical aspects of care, such as communication between providers, prevention of, and response to complications, patient safety and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This complication measure was developed to identify institutions, whose performance is better or worse than expected based on their patient case mix, and therefore promote hospital quality improvement and better inform consumers about the quality of care.

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

N/A. This measure is not an intermediate outcome, process, or structure performance measure.

****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.

In 2010, there were 168,000 THAs and 385,000 TKAs performed on Medicare beneficiaries 65 years and older (National Center for Health Statistics, 2010). There is an increasing trend in both of these procedures, with some projecting that annual TKA and THA volume will reach more than 3 million and 500,000 by 2030 respectively (Kurtz et al., 2007; Kurtz et al., 2014). Although these procedures dramatically improve quality of life, they are costly. In 2005, annual hospital charges totaled $3.95 billion and $7.42 billion for primary THA and TKA, respectively (Kurtz et al., 2007). These costs are projected to increase significantly for both THAs and TKAs by 2020 (Kurtz et al., 2014). Medicare is the single largest payer for these procedures, covering approximately two-thirds of all THAs and TKAs performed in the US (Ong et al., 2006). Combined, THA and TKA procedures account for the largest procedural cost in the Medicare budget (Bozic et al., 2008).

Since THAs and TKAs are commonly performed and costly procedures, it is imperative to address quality of care. Complications increase costs associated with THA and TKA and affect the quality, and potentially quantity, of life for patients. Although complications following elective THA and TKA are rare, the results can be devastating. Rates for periprosthetic joint infection following THA and TKA range from 1.6% to 2.3%, depending upon the population (Bongartz et al., 2008; Kurtz et al., 2010). Reported 90-day death rates following THA range from 0.7% (Soohoo et al., 2010) to 2.7% (Cram et al., 2007). Rates for pulmonary embolism following TKA range from 0.5% to 0.9% (Cram et al., 2007; Mahomed et al., 2003; Khatod et al., 2008; Solomon et al., 2006; Bozic et al., 2014). Rates for wound infection in Medicare population-based studies vary between 0.3% and 1.0% (Cram et al., 2007; Mahomed et al., 2003; Solomon et al., 2006; Bozic et al., 2014). Rates for septicemia range from 0.1%, during the index admission (Browne et al., 2010) to 0.3%, 90 days following discharge for primary TKA (Cram et al., 2007; Bozic et al., 2014). Rates for bleeding and hematoma following TKA range from 0.9% (Browne et al., 2010; Bozic et al., 2014) to 1.7% (Huddleston et al., 2009).

The variation in complication rates across hospitals indicates there is room for quality improvement and targeted efforts to reduce these complications could result in better patient care and potential cost savings (Navathe et al, 2017; Cyriac et al., 2016; Borza et al., 2019). Measurement of patient outcomes allows for a comprehensive view of quality of care that reflects complex aspects of care such as communication between providers and coordinated transitions to the outpatient environment. These aspects are critical to patient outcomes, and are broader than what can be captured by individual process of care measures.

The THA/TKA hospital-specific risk-standardized complication rate (RSCR) measure is thus intended to inform quality-of-care improvement efforts, as individual process-based performance measures cannot encompass all the complex and critical aspects of care within a hospital that contribute to patient outcomes.

References:

Bongartz T, Halligan CS, Osmon DR, et al. Incidence and risk factors of prosthetic joint infection after total hip or knee replacement in patients with rheumatoid arthritis. Arthritis Rheum. Dec 15 2008;59(12):1713-1720.

Borza T, Oerline MK, Skolarus TA, et al. Association Between Hospital Participation in Medicare Shared Savings Program Accountable Care Organizations and Readmission Following Major Surgery. *Ann Surg*. 2019;269(5):873-878. doi:10.1097/SLA.0000000000002737.

Bozic KJ, Grosso LM, Lin Z, et al. Variation in hospital-level risk-standardized complication rates following elective primary total hip and knee arthroplasty. *J Bone Joint Surg Am*. 2014;96(8):640-647. doi:10.2106/JBJS.L.01639.

Bozic KJ, Rubash HE, Sculco TP, Berry DJ. An analysis of medicare payment policy for total joint arthroplasty. *J Arthroplasty.* Sep 2008;23(6 Suppl 1):133-138.

Browne J, Cook C, Hofmann A, Bolognesi M. Postoperative morbidity and mortality following total knee arthroplasty with computer navigation. Knee. Mar 2010;17(2):152-156.

Cram P, Vaughan-Sarrazin MS, Wolf B, Katz JN, Rosenthal GE. A comparison of total hip and knee replacement in specialty and general hospitals. J Bone Joint Surg Am. Aug 2007;89(8):1675-1684.

Cyriac, James MD; Garson, Leslie MD; Schwarzkopf, Ran MD; Ahn, Kyle MD; Rinehart, Joseph MD; Vakharia, Shermeen MD, MBA; Cannesson, Maxime MD, PhD; Kain, Zeev MD, MBA. Total Joint Replacement Perioperative Surgical Home Program: 2-Year Follow-Up, Anesthesia & Analgesia: July 2016 - Volume 123 - Issue 1 - p 51-62 doi: 10.1213/ANE.0000000000001308.

Huddleston JI, Maloney WJ, Wang Y, Verzier N, Hunt DR, Herndon JH. Adverse Events After Total Knee Arthroplasty: A National Medicare Study. The Journal of Arthroplasty. 2009;24(6, Supplement 1):95-100.

Khatod M, Inacio M, Paxton EW, et al. Knee replacement: epidemiology, outcomes, and trends in Southern California: 17,080 replacements from 1995 through 2004. Acta Orthop. Dec 2008;79(6):812-819.

Kurtz S, Ong K, Lau E, Bozic K. Impact of the economic downturn on total joint replacement demand in the United States: updated projections to 2021. J Bone Joint Surg Am, 96 (2014), pp. 624-630.

Kurtz S, Ong K, Lau E, Bozic K, Berry D, Parvizi J. Prosthetic joint infection risk after TKA in the Medicare population. Clin Orthop Relat Res. 2010;468:5.

Kurtz S, Ong K, Lau E, Mowat F, Halpern M. Projections of primary and revision hip and knee arthroplasty in the United States from 2005 to 2030. J Bone Joint Surg Am. 2007 Apr;89(4):780-5.

Mahomed NN, Barrett JA, Katz JN, et al. Rates and outcomes of primary and revision total hip replacement in the United States medicare population. J Bone Joint Surg Am. Jan 2003;85-A(1):27-32.

National Center for Health Statistics. National Hospital Discharge Survey: 2010 table, Procedures by selected patient characteristics - Number by procedure category and age. Available at http://www.cdc.gov/nchs/data/nhds/4procedures/2010pro4_numberprocedureage.pdf.

Navathe AS, Troxel AB, Liao JM, et al. Cost of Joint Replacement Using Bundled Payment Models. *JAMA Intern Med.* 2017;177(2):214–222. doi:10.1001/jamainternmed.2016.8263.

Ong KL, Mowat FS, Chan N, Lau E, Halpern MT, Kurtz SM. Economic burden of revision hip and knee arthroplasty in Medicare enrollees. Clin Orthop Relat Res. May 2006;446:22-28.

Solomon DH, Chibnik LB, Losina E, et al. Development of a preliminary index that predicts adverse events after total knee replacement. Arthritis & Rheumatism. 2006;54(5):1536-1542.

Soohoo NF, Farng E, Lieberman JR, Chambers L, Zingmond DS. Factors That Predict Short-term Complication Rates After Total Hip Arthroplasty. Clin Orthop Relat Res. Sep 2010;468(9):2363-2371.

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

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

☐ Clinical Practice Guideline recommendation (with evidence review)

☐ US Preventive Services Task Force Recommendation

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

☐ Other

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

N/A

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

N/A

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

N/A

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

N/A

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.

The goal of this measure is to improve patient outcomes by providing patients, physicians, hospitals, and policy makers with information about hospital-level, risk-standardized complication rates (RSCRs) following hospitalization for primary elective THA and TKA. Measurement of patient outcomes allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. Complex and critical aspects of care, such as communication between providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment, all contribute to patient outcomes but are difficult to measure by individual process measures. The goal of outcomes measurement is to risk adjust for patients' conditions at the time of hospital admission and then evaluate patient outcomes. This measure was developed to identify institutions whose performance is better or worse than would be expected based on each institution's patient case mix, and therefore promote hospital quality improvement and better inform consumers about care quality.

THA and TKA complications is a priority area for outcome measure development, as it is an outcome that is likely attributable to care processes and is an important outcome for patients. Measuring and reporting complication rates will inform healthcare providers and facilities about opportunities to improve care, strengthen incentives for quality improvement, and ultimately improve the quality of care received by Medicare patients. The measure will also provide patients with information that could guide their choices. In addition, it has the potential to lower health care costs associated with complications.

1b.2. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis. *(This is required for 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.*

Variation in complication rates indicates opportunity for improvement. We conducted analyses using data from April 1, 2016 to March 31, 2019 Medicare administrative claims data (n= 962,744 admissions from 3,418 hospitals).

The three-year hospital-level risk standardized complication rate (RSCR) has a mean of 2.5% and range from 1.2-10.6% in the study cohort. As shown below, the median risk-standardized rate is 2.4%. The distribution of RSCRs across hospitals is shown below:

Distribution of Hospital THA/TKA RSCRs over Different Time Periods

Results for each data year

Characteristic//04/2016-03/2017//04/2017-03/2018//04/2018-03/2019//04/2016-03/2019

Number of Hospitals//3274//3271//3250//3418

Number of Admissions//336445//330765//295534//962744

Mean (SD)//2.6(0.4)//2.4(0.4)//2.3(0.3)//2.5(0.5)

Range (Min-Max)//1.1-9.3//1.3-13//1.2-4.5//1.2-10.6

Minimum//1.1//1.3//1.2//1.2

10th percentile//2.1//2.1//2.0//1.9

20th percentile//2.3//2.2//2.1//2.1

30th percentile//2.4//2.3//2.2//2.3

40th percentile//2.5//2.3//2.2//2.3

50th percentile//2.5//2.4//2.3//2.4

60th percentile//2.5//2.4//2.3//2.5

70th percentile//2.7//2.5//2.4//2.6

80th percentile//2.8//2.7//2.5//2.8

90th percentile//3.0//2.9//2.7//3.0

Maximum//9.3//13.0//4.5//10.6

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

N/A

1b.4. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability. *(This is required for 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.*

Distribution of THA/TKA RSCRs by Proportion of Dual Eligible Patients:

Data Source: Medicare FFS claim and Medicare Beneficiary Summary File (MBSF) data

Dates of Data: April 1, 2016 to March 31, 2019

Variation in RSCRs across hospitals (with at least 25 cases) by proportion of patients with social risk//

Description of Social Risk Variable//Dual Eligibility

Quartile//Q1//Q4

Social Risk Proportion (%)//(0-5.32)//(21.26-79.17)

# of Hospitals//690//691

100%Max//4.1//6.0

90%//3.0//3.2

75%//2.6//2.8

50%//2.3//2.5

25%//2.1//2.2

10%//1.8//2.1

0%Min//1.2//1.5

Distribution of THA/TKA RSCRs by Proportion of Patients with AHRQ SES Index Scores:

Data Source: Medicare FFS claims and The American Community Survey (2013-2017) data

Dates of Data: April 1, 2016 to March 31, 2019

Variation in RSCRs across hospitals (with at least 25 cases) by proportion of patients in lower and upper social risk quartiles//

Description of Social Risk Variable //AHRQ SES Index

Quartile//Q1//Q4

Social Risk Proportion (%)//(0-1.74)//(6.82-95.38)

# of Hospitals//690//691

100%Max//4.6//6.9

90%//3.0//3.1

75%//2.7//2.8

50%//2.3//2.5

25%//2.1//2.2

10%//1.8//2.1

0%Min//1.3//1.6

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

N/A

2. Reliability and Validity—Scientific Acceptability of Measure Properties

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

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

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

Musculoskeletal : Joint Surgery, Musculoskeletal : Osteoporosis, Musculoskeletal : Rheumatoid Arthritis, Surgery

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

Care Coordination, Safety, Safety : Complications, Safety : Overuse

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

Elderly, Populations at Risk

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

https://qualitynet.org/inpatient/measures/complication/methodology

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

This is not an eMeasure **Attachment:**

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

Attachment**:** NQF_datadictionary_THATKAcomp_Fall2020_final_7.22.20.xlsx

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

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

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

Not an instrument-based measure

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

No

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

Updates consisted of updating the specifications to include new and modified ICD-10 CM/PCS codes.

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

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

The outcome for this measure is any complication occurring during the index admission (not coded present on arrival) to 90 days post-date of the index admission. Complications are counted in the measure only if they occur during the index hospital admission or during a readmission. The complication outcome is a dichotomous (yes/no) outcome. If a patient experiences one or more of these complications in the applicable time period, the complication outcome for that patient is counted in the measure as a “yes”.

S.5. 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).*

The composite complication is a dichotomous outcome (yes for any complication(s); no for no complications). Therefore, if a patient experiences one or more complications, the outcome variable will get coded as a "yes". Complications are counted in the measure only if they occur during the index hospital admission (and are not present on admission) or during a readmission.

The complications captured in the numerator are identified during the index admission OR associated with a readmission up to 90 days post-date of index admission, depending on the complication. The follow-up period for complications from date of index admission is as follows:

The follow-up period for AMI, pneumonia, and sepsis/septicemia/shock is seven days from the date of index admission because these conditions are more likely to be attributable to the procedure if they occur within the first week after the procedure. Additionally, analyses indicated a sharp decrease in the rate of these complications after seven days.

Death, surgical site bleeding, and pulmonary embolism are followed for 30 days following admission because clinical experts agree these complications are still likely attributable to the hospital performing the procedure during this period and rates for these complications remained elevated until roughly 30 days post admission.

The measure follow-up period is 90 days after admission for mechanical complications and periprosthetic joint infection/wound infection. Experts agree that mechanical complications and periprosthetic joint infection/wound infections due to the index THA/TKA occur up to 90 days following THA/TKA.

The measure counts all complications occurring during the index admission regardless of when they occur. For example, if a patient experiences an AMI on day 10 of the index admission, the measure will count the AMI as a complication, although the specified follow-up period for AMI is seven days. Clinical experts agree with this approach, as such complications likely represent the quality of care provided during the index admission.

As of 2014 reporting, the measure does not count complications in the complications outcome that are coded as present on admission (POA) during the index admission; this prevents identifying a condition as a complication of care if it was present on admission for the THA/TKA procedure.

For full list of codes defining complications, see the Data Dictionary attached in field S.2b.

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

The target population for the publicly reported measure includes admissions for Medicare FFS beneficiaries who are at least 65 years of age undergoing elective primary THA and/or TKA procedures.

Additional details are provided in S.7 Denominator Details.

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).*

To be included in the measure cohort used in public reporting, patients must meet the following additional inclusion criteria:

1. Enrolled in Medicare fee-for-service (FFS) Part A and Part B for the 12 months prior to the date of admission; and enrolled in Part A during the index admission;
2. Aged 65 or older
3. Having a qualifying elective primary THA/TKA procedure; elective primary THA/TKA procedures are defined as those procedures without any of the following:
    - Fracture of the pelvis or lower limbs coded in the principal or secondary discharge diagnosis fields on the index admission claim (Note: Periprosthetic fractures must be additionally coded as present on admission [POA] in order to disqualify a THA/TKA from cohort inclusion, unless exempt from POA reporting.);
    - A concurrent partial hip or knee arthroplasty procedure;
    - A concurrent revision, resurfacing, or implanted device/prosthesis removal procedure;
    - Mechanical complication coded in the principal discharge diagnosis field on the index admission claim;

- Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field on the index admission claim; or,
- Transfer from another acute care facility for the THA/TKA.

Patients are eligible for inclusion in the denominator if they had an elective primary THA and/or a TKA AND had continuous enrollment in Part A and Part B Medicare fee-for-service (FFS) 12 months prior to the date of index admission.

This measure can also be used for an all-payer population aged 18 years and older. We have explicitly tested the measure in both patients aged 18+ years and those aged 65+ years (see Testing Attachment for details).

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

This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare;
2. Who were discharged against medical advice (AMA); or,
3. Who had more than two THA/TKA procedure codes during the index hospitalization.

After applying these exclusion criteria, we randomly select one index admission for patients with multiple index admissions in a calendar year. We therefore exclude the other eligible index admissions in that year.

S.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.)*

This measure excludes index admissions for patients:

1. Without at least 90 days post-discharge enrollment in FFS Medicare

Rationale: The 90-day complication outcome cannot be assessed in this group since claims data are used to determine whether a complication of care occurred.

2. Who were discharged against medical advice (AMA); or,

Rationale: Providers did not have the opportunity to deliver full care and prepare the patient for discharge.

3. Who had more than two THA/TKA procedure codes during the index hospitalization

Rationale: Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

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

N/A

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

Statistical risk model

If other:

S.12. Type of score:

Rate/proportion

If other:

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

Better quality = Lower score

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

The measure estimates hospital-level RSCRs following elective primary THA/TKA using hierarchical logistic regression models. In brief, the approach simultaneously models data at the patient and hospital levels to account for variance in patient outcomes within and between hospitals (Normand and Shahian, 2007). At the patient level, it models the log-odds of a complication occurring within 90 days of the index admission using age, sex, selected clinical covariates, and a hospital-specific intercept. At the hospital level, it models the hospital-specific intercepts as arising from a normal distribution. The hospital intercept represents the underlying risk of a complication at the hospital, after accounting for patient risk. The hospital-specific intercepts are given a distribution to account for the clustering (non-independence) of patients within the same hospital. If there were no differences among hospitals, then after adjusting for patient risk, the hospital intercepts should be identical across all hospitals.

The RSCR is calculated as the ratio of the number of "predicted" to the number of "expected" admissions with a complication at a given hospital, multiplied by the national observed complication rate. For each hospital, the numerator of the ratio is the number of complications within 90 days predicted on the basis of the hospital's performance with its observed case mix, and the denominator is the number of complications expected based on the nation's performance with that hospital's case mix. This approach is analogous to a ratio of "observed" to "expected" used in other types of statistical analyses. It conceptually allows for a comparison of a particular hospital's performance given its case mix to an average hospital's performance with the same case mix. Thus, a lower ratio indicates lower-than-expected complication rates or better quality, and a higher ratio indicates higher-than-expected complication rates or worse quality.

The "predicted" number of admissions with a complication (the numerator) is calculated by using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of having an admission with a complication. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are log transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of admissions with a complication (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific effect. The results are log transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimate the model coefficients using the years of data in that period.

This calculation transforms the ratio of predicted over expected into a rate that is compared to the national observed complication rate. The hierarchical logistic regression models are described fully in the original methodology report posted on QualityNet:
https://www.qualitynet.org/inpatient/measures/complication/methodology.

References:

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

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

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

N/A

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

If other, please describe in S.18.

Claims, Enrollment Data

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

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

Data sources for the Medicare FFS measure:

Medicare Part A Inpatient and Part B Outpatient Claims: This data source contains claims data for FFS inpatient and outpatient services including Medicare inpatient hospital care, outpatient hospital services, as well as inpatient and outpatient physician claims for the 12 months prior to an index admission.

Medicare Enrollment Database (EDB): This database contains Medicare beneficiary demographic, benefit/coverage, and vital status information. This data source was used to obtain information on several inclusion/exclusion indicators such as Medicare status on admission as well as vital status. These data have previously been shown to accurately reflect patient vital status (Fleming et al., 1992). The Master Beneficiary Summary File (MBSF) is an annually created file derived the EDB that contains enrollment information for all Medicare beneficiaries including dual eligible status. Years 2016-2019 were used.

The American Community Survey (2013-2017): We used the American Community Survey (2013-2017) to derive an updated Agency for Healthcare Research and Quality (AHRQ) Socioeconomic (SES) index score at the patient nine-digit zip code level for use in studying the association between our measure and social risk factors (SRFs).

References:

Fleming C., Fisher ES, Chang CH, Bubolz D, Malenda J. Studying outcomes and hospital utilization in the elderly: The advantages of a merged data base for Medicare and Veterans Affairs Hospitals. Medical Care. 1992; 30(5): 377-91.

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

No data collection instrument provided

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

Facility

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

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.)*

N/A

3. Validity – See attached Measure Testing Submission Form

NQF_testing_THATKAcomplications_Fall2020_final_11.02.20-637418270788250049.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.

Yes

2.2 For maintenance of endorsement

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

Yes

2.3 ***For maintenance of endorsement***

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

Yes - Updated information is included

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

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

Measure Number (*if previously endorsed*)**:** 1550
Measure Title: Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
Date of Submission: **11/3/2020**

Type of Measure:

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

*cell intentionally left blank

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

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

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

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.17*)	Measure Tested with Data From:
☐ abstracted from paper record	☐ abstracted from paper record
☒ claims	☒ claims
☐ registry	☐ registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☒ other: Medicare Enrollment Data (including the Master Beneficiary Summary File)	☒ other: Census Data/American Community Survey, Medicare Enrollment Data (including the Master Beneficiary Summary File)

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

The datasets used for testing included Medicare Parts A and B claims, as well as the Medicare Enrollment Database (EDB). Additionally, the American Community Survey census data as well as enrollment data were used to assess socioeconomic factors (dual eligible variable obtained through enrollment data; Agency for Healthcare Research and Quality [AHRQ] socioeconomic status [SES] index obtained through census data). The dataset used varies by testing type; see Section 1.7 for details.

1.3. What are the dates of the data used in testing? The dates used vary by testing type; see Section 1.7 for details.

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

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item 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: Click here to describe	☐ other: Click here to describe

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

For this measure, hospitals are the measured entities. All non-federal, short-term acute care inpatient US hospitals (including territories) with Medicare fee-for-service (FFS) beneficiaries aged 65 years and older are included. The number of measured entities (hospitals) varies by testing type; see Section 1.7 for details.

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

The number of admissions/patients varies by testing type: see Section 1.7 for details.

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

The datasets, dates, number of measured hospitals, and number of admissions used in each type of testing are in Table 1.

Measure Development

For measure development, we used Medicare administrative claims data (2008). The dataset also included administrative data on each patient for the 12 months prior to the index admission and the 90 days following it. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data. We randomly split the data (2008) into two equal samples: **the Development Dataset** and **Internal Validation Dataset.**

Measure Testing

For analytical updates for this measure, we used three-years of Medicare administrative claims data (April 2016 – March 2019). The dataset also included administrative data on each patient for the 12 months prior to the index admission and the 90 days following it. The dataset contained inpatient and facility outpatient claims and Medicare enrollment database (EDB) data.

Table 1. Dataset Descriptions

Dataset	Applicable Section in the Testing Attachment	Description of Dataset
Development and Validation Datasets (Medicare Fee-For-Service Administrative Claims Data)	Section 2b3 Risk Adjustment/Stratification **2b3.6. Statistical Risk Model Discrimination Statistics** **2b3.7. Statistical Risk Model Calibration Statistics**	Entire Cohort: Dates of Data: 2008 Number of admissions = 290,329 Patient Descriptive Characteristics: mean age = 75.2 years; % male = 35.7% Number of measured hospitals: 3,223 This cohort was randomly split for initial model testing. First half of split sample -Number of Admissions: 145,206 -Number of Measured Hospitals: 3,221 Second half of split sample -Number of Admissions: 145,123 -Number of Measured Hospitals: 3,223

Dataset	Applicable Section in the Testing Attachment	Description of Dataset
Testing Dataset Medicare Fee-For-Service Administrative Claims Data (April 1, 2016 – March 30, 2019)	Section 2a2 Reliability Testing Section 2b1 Validity Testing Section 2b2 Testing of Measure Exclusion Section 2b3 Risk Adjustment/Stratification **2b3.6. Statistical Risk Model Discrimination Statistics** Section 2b4 Meaningful Differences	Dates of Data: April 2016 – March 2019 Number of admissions = 962,744 Patient Descriptive Characteristics: meanage = 73.9 years; % male = 37.2 Number of measured hospitals: 3,418
The American Community Survey (ACS)	Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures	Dates of Data: 2013-2017 We used the AHRQ SES index score derived from the American Community Survey (2013-2017) to study the association between the 90-day complication outcome and SRFs. The AHRQ SES index score is based on beneficiary 9-digit zip code level of residence and incorporates 7 census variables found in the American Community Survey.
Master Beneficiary Summary File (MBSF)	Section 2b3: Risk adjustment/Stratification for Outcome or Resource Use Measures	Dates of Data: April 2016 – March 2019 We used dual eligible status (for Medicare and Medicaid) derived from the MBSF to study the association between the 90-day measure outcome and dual-eligible status.

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.

We selected social risk factor (SRF) variables to analyze after reviewing the literature and examining available national data sources. We sought to find variables that are consistently captured in a reliable fashion for all patients in this measure. There is a large body of literature linking various SRFs to worse health status and higher complications over a lifetime. Income, education, and occupation are the most commonly examined SRFs studied. The causal pathways for SRF variable selection are described below in Section 2b3.3a. Unfortunately these variables are not available at the patient level for this measure. Therefore proxy measures of income, education level and economic status were selected.

The SRF variables used for analysis were:

- Dual eligible status: Dual eligible status (in other words, being enrolled in both Medicare and Medicaid) patient-level data is obtained from the CMS Master Beneficiary Summary File (MBSF)

Following guidance from ASPE and a body of literature demonstrating differential health care and health outcomes among dual eligible patients, we identified dual eligibility as a key variable (ASPE 2016; ASPE 2020). We recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patients' income or assets because it does not provide a range of results and is only a dichotomous outcome. However, the threshold for over 65-year-old Medicare patients is valuable, as it takes into account both income and assets and is consistently applied across states for the older population. We acknowledge that it is important to test a wider variety of SRFs including key variables such as education and poverty level; therefore, we also tested a validated composite based on census data linked to as small a geographic unit as possible.

- AHRQ-validated SES index score (summarizing the information from the following 7 variables): percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room)

Finally, we selected the AHRQ SES index score because it is a well-validated variable that describes the average SES of people living in defined geographic areas (Bonito et al., 2008). Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. We considered the area deprivation index (ADI) among many other potential indicators when we initially evaluated the impact of SDS indicators. We ultimately did not include the ADI at the time, partly due to the fact that the coefficients used to derive ADI had not been updated for many years. Recently, the coefficients for ADI have been updated and therefore we compared the ADI with the AHRQ SES Index and found them to be highly correlated. In this submission, we present analyses using the census block level, the most granular level possible using American Community Survey (ACS) data. A census block group is a geographical unit used by the US Census Bureau which is between the census tract and the census block. It is the smallest geographical unit for which the bureau publishes sample data. The target size for block groups is 1,500 and they typically have a population of 600 to 3,000 people. We used 2013-2017 ACS data and mapped patients' 9-digit ZIP codes via vendor software to the census block group level. Given the variation in cost of living across the country, the median income and median property value components of the AHRQ SES Index were adjusted by regional price parity values published by the Bureau of Economic Analysis (BEA). This provides a better marker of low SES neighborhoods in high expense geographic areas. We then calculated an AHRQ SES Index score for census block groups that can be linked to 9-digit ZIP codes. We used the percentage of patients with an AHRQ SES index score equal to or below 42.7 to define the lowest quartile of the AHRQ SES Index.

References:

Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. Health affairs (Project Hope). 2002; 21(2):60-76.

Bonito A, Bann C, Eicheldinger C, Carpenter L. Creation of new race-ethnicity codes and socioeconomic status (SES) indicators for Medicare beneficiaries. Final Report, Sub-Task. 2008;2.

Courtney M, Huddleston J, Iorio R, Markel D. Socioeconomic Risk Adjustment Models for Reimbursement Are Necessary in Primary Total Joint Arthroplasty. July 2016; 32(1):1-5. https://doi.org/10.1016/j.arth.2016.06.050.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Report to Congress: Social Risk factors and Performance Under Medicare's Value-based Payment Programs. 2016; https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs. Accessed November 10, 2019.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

Martsolf G, Barrett M, Weiss A, Kandrack R, Washington R, Steiner C, Mehrotra A, SooHoo N, Coffey R. Impact of Race/Ethnicity and Socioeconomic Status on Risk-Adjusted Hospital Readmission Rates Following Hip and Knee Arthroplasty, The Journal of Bone and Joint Surgery. 2016;98(16):1385-1391. https://doi.org/10.2106/JBJS.15.00884.

White, R.S., Sastow, D.L., Gaber-Baylis, L.K. *et al.* Readmission Rates and Diagnoses Following Total Hip Replacement in Relation to Insurance Payer Status, Race and Ethnicity, and Income Status. *J. Racial and Ethnic Health Disparities* 5, 1202–1214 (2018). https://doi.org/10.1007/s40615-018-0467-0.

Xu HF, White RS, Sastow DL, Andreae MH, Gaber-Baylis LK, Turnbull ZA. Medicaid insurance as primary payer predicts increased mortality after total hip replacement in the state inpatient databases of California, Florida and New York. *J Clin Anesth*. 2017;43:24-32. doi:10.1016/j.jclinane.2017.09.008.

2a2. RELIABILITY TESTING

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

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

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

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

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

Measure Score Reliability

We performed two types of reliability testing. First, we estimated the overall measure score reliability by calculating the intra-class correlation coefficient (ICC) using a split sample (in other words, test-retest) method. Second, we estimated the facility-level reliability (signal-to-noise reliability).

Split-Sample Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is naturally the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. Accordingly, our approach to assessing reliability is to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. That is, we take a "test-retest" approach in which hospital performance is measured once using a random subset of patients, and then measured again using a second random subset exclusive of the first, and the agreement of the two resulting performance measures compared across hospitals (Rousson, Gasser, and Seifert, 2002).

For split-sample reliability of the measure in aged 65 years and older, we randomly sampled half of patients within each hospital for a three year period, calculated the measure for each hospital, and repeated the calculation using the second half. Thus, each hospital is measured twice, but each measurement is made using an entirely distinct set of patients. To the extent that the calculated measures of these two subsets agree, we have evidence that the measure is assessing an attribute of the hospital, not of the patients. As a metric of agreement we calculated the intra-class correlation coefficient (Shrout & Fleiss, 1979), and assessed the values according to conventional standards (Landis & Koch, 1977). Specifically, we used a combined 2016-2019 sample, randomly split it into two approximately equal subsets of patients, and calculated the RSCR for each hospital for each sample. The agreement of the two RSCRs was quantified for hospitals in each sample using the intra-class correlation as defined by ICC (2,1). (Shrout & Fleiss, 1979)

Using two non-overlapping random samples provides a conservative estimate of the measure's reliability, compared with using two random but potentially overlapping samples which would exaggerate the agreement. Moreover, because our final measure is derived using hierarchical logistic regression, and a known property of hierarchical logistic regression models is that smaller volume hospitals contribute less 'signal', a split sample using a single measurement period would introduce extra noise. This leads to an underestimate in the actual test-retest reliability that would be achieved if the measure were reported using the full measurement period, as evidenced by the Spearman Brown prophecy formula (Spearman 1910, Brown 1910). We used this formula to estimate the reliability of the measure if the whole cohort were used, based on an estimate from half the cohort.

Signal-to-Noise

We estimated the signal to noise reliability (facility-level reliability), which is the reliability with which individual units (hospitals) are measured. While test re-test reliability is the most relevant metric from the perspective of overall measure reliability, it is also meaningful to consider the separate notion of "unit" reliability, that is, the reliability with which individual units (here, hospitals) are measured. The reliability of any one facility's measure score will vary depending on the number of patients admitted for an elective THA/TKA procedure. Facilities with more volume (in other words, with more patients) will tend to have more reliable scores, while facilities with less volume will tend to have less reliable scores. Therefore, we used the formula presented by Adams and colleagues (2010) to calculate facility-level reliability.

Where facility-to-facility variance is estimated from the hierarchical logistic regression model, n is equal to each facility's observed case size, and the facility error variance is estimated using the variance of the logistic distribution ($\pi^2/3$). The facility-level reliability testing is limited to facilities with at least 25 admissions for public reporting.

Signal to noise reliability scores can range from 0 to 1. A reliability of zero implies that all the variability in a measure is attributable to measurement error. A reliability of one implies that all the variability is attributable to real difference in performance.

Additional Information

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

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

Furthermore, we assessed the variation in the frequency of the variables over time: Detailed information is presented in the measure's 2020 Condition-Specific Measure Updates and Specifications Report cited below.

References:

Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

DeBuhr J, McDowell, K, Grady J, et al., 2020 Condition-Specific Complication Measure Updates and Specifications Report - Available at: https://www.qualitynet.org/inpatient/measures/complication/methodology.

Landis J, Koch G, The measurement of observer agreement for categorical data, Biometrics, 1977;33:159-174.

Rousson V, Gasser T, Seifert B. "Assessing intrarater, interrater and test–retest reliability of continuous measurements," Statistics in Medicine, 2002, 21:3431-3446.

Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 1979, 86, 420-3428.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. British Journal of Psychology, 3, 271–295.

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

Measure Score Reliability Results

Signal-to-Noise

We calculated the signal-to-noise reliability score for each hospital with at least 25 admissions* (see Table 2 below). The median reliability score was 0.87, ranging from 0.46 to 1.00. The 25th and 75th percentiles were 0.74 and 0.94, respectively. The median reliability score demonstrates moderate reliability.

Table 2. Signal-to-noise reliability distribution for THA/TKA complications

Mean	Std. Dev.	Min	5th Percentile	10th Percentile	25th Percentile	Median	75th Percentile	90th Percentile	95th Percentile	Max
0.83	0.14	0.46	0.53	0.60	0.74	0.87	0.94	0.97	0.98	1.00

*Hospital measure scores are calculated for all hospitals (including those that have fewer than 25 procedures) but only publicly reported for those that have at least 25 procedures to ensure hospital results are reliable.

Split-Sample Reliability

In total, 962,744 admissions were included in the analysis, using 3 years of data. After randomly splitting the sample into two halves, there were 480,496 admissions from 3,365 hospitals in one half and 482,248 admissions from 3,418 hospitals in the other half. As a metric of agreement, we calculated the ICC for hospitals with 25 admissions or more. Using the Spearman-Brown prediction formula, the agreement between the two independent assessments of the RSCR for each hospital was 0.524.

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?*)

Measure Score Reliability Results

Using the approach used by Adams et. al. and Yu et al., we obtained the median signal-to-noise reliability score of 0.87, which demonstrates “almost perfect” agreement.

Our interpretation of the results is based on the standards established by Landis and Koch (1977):

< 0 – Less than chance agreement;

0 – 0.2 Slight agreement;

0.21 – 0.39 Fair agreement;

0.4 – 0.59 Moderate agreement;

0.6 – 0.79 Substantial agreement;

0.8 – 0.99 Almost Perfect agreement; and

1 Perfect agreement

The split-sample reliability score of 0.524, discussed in the previous section, represents the lower bound of estimate of the true measure reliability.

In the absence of empirically supported standards, our position is that 'acceptability' depends on context. For simple concepts or constructs, such as a patient's weight, the expectation is that the test-retest reliability of a measure of that construct should be quite high. However, for complex constructs, such as clinical severity, patient comorbidity, or symptom profiles used to identify a condition or clinical state, reliability of measures used to define these constructs is quite a bit lower.

Taken together, these results indicate that there is substantial reliability in the measure score.

References:

Adams J, Mehrota, A, Thoman J, McGlynn, E. (2010). Physician cost profiling – reliability and risk of misclassification. NEJM, 362(11): 1014-1021.

Landis J, Koch G. The measurement of observer agreement for categorical data, Biometrics 1977;33:159-174.

Yu, H, Mehrota, A, Adams J. (2013). Reliability of utilization measures for primary care physician profiling. Healthcare, 1, 22-29.

2b1. VALIDITY TESTING

2b1.1. What level of validity testing was conducted? (*may be one or both levels*)

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

☒ **Performance measure score**

☒ **Empirical validity testing**

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

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

Empirical Validity

Stewards of NQF-endorsed measures going through the re-endorsement process are required to demonstrate external validity testing at the time of maintenance review, or if this is not possible, justify the use of face validity only. To meet this requirement for the THA/TKA complications measure, we identified and assessed the measure's correlation with other measures that target the same domain of quality (e.g. complications, safety, or post-procedure utilization) for the same or similar populations. The goal was to identify if better performance in this measure was related to better performance on other relevant structural or outcomes measures. After literature review and consultations with measures experts in the field, there were very few measures identified that assess the same domains of quality. Given that challenge, we selected the following to use for validity testing:

1. Overall Hospital Star Rating: CMS's Overall Hospital Star Rating assesses hospitals' overall performance (expressed on Hospital Compare graphically, as stars) based on a weighted average of "group scores" from different domains of quality (mortality, readmissions, safety, patient experience, imaging, effectiveness of care, timeliness of care). Each group has within it, measures that are reported on Hospital Compare, including this complications measure. Group scores for each individual group are derived from latent-variable models that identify an underlying quality trait for each group. Group scores are combined into an overall hospital score using fixed weights; overall hospital scores are then clustered, using k-means clustering, into five groups and are assigned one-to-five stars (the hospital's Star Rating). For the validity testing presented in this testing form, we used hospital's Star Ratings from 3,418 Medicare FFS hospitals from March 2019. The full methodology for the Overall Hospital Star

Rating can be found at: https://www.qualitynet.org/outpatient/public-reporting/overall-ratings/resources.

2. Hospital THA/TKA Surgical Volume: There is evidence that surgical complication rates for providers (both surgeons and hospitals) decline with increasing volume (Sibley et al., 2017; Murphy et al., 2019; Courtney et al., 2018). Thus, we assessed validity of the measure by examining the relationship between volume and the measure score for hospitals. To establish validity, we expect scores to be correlated with case volume at the hospital level.

We examined the relationship of performance between the THA/TKA complication measure scores (RSCRs) and each of these external measures of hospital quality. For the external measures, the comparison was against performance within quartiles of the Star Ratings overall category (1-5 Stars), as well as hospital THA/TKA surgical volume. We predicted the THA/TKA complication measure scores would have a small association with the overall hospital star rating scores, with lower RSCRs associated with better Star ratings. With THA/TKA surgical volume, we assume that lower RSCRs will be moderately associated with higher volume hospitals.

References

Courtney M, Frisch N, Bohl D, Della Valle C. Improving Value in Total Hip and Knee Arthroplasty: The Role of High Volume Hospitals. *The Journal of Arthroplasty.* 2018;33(1):1-5. https://doi.org/10.1016/j.arth.2017.07.040.

Murphy WS, Cheng T, Lin B, Terry D, Murphy SB. Higher Volume Surgeons Have Lower Medicare Payments, Readmissions, and Mortality After THA. *Clin Orthop Relat Res*. 2019;477(2):334-341. doi:10.1097/CORR.0000000000000370.

Sibley R, Charumbhumi, V, Hutzler L, Paoli A, Bosco, J. Joint Replacement Volume Positively Correlates With Improved Hospital Performance on Centers for Medicare and Medicaid Services Quality Metrics. *The Journal of Arthroplasty*. 2017;32(5):1409-1413. https://doi.org/10.1016/j.arth.2016.12.010.

Empirical Validation of Claims-Based Definition of Complications

During original measure development we validated the administrative claims-based definition of THA/TKA complication (original model specification) against medical record data (Dataset 2). The primary goal of this validation study was to determine the overall agreement between patients identified as having a complication (or no complication) using claims data compared with those who had a complication (or no complication) documented in the medical record. We conducted a secondary analysis of agreement of individual, specific complications to identify opportunities for measure improvement.

A statistician and practicing rheumatologist conducted a detailed analysis of each abstracted patient record and compared the findings to the patient results found in the claims data. If any disagreement between the medical record abstraction and the claims data was found, the disagreement was documented and explored in further detail. In some instances, we requested that the medical record be re-abstracted in order to confirm the disagreement and/or to obtain more clinical information. Our clinical team also reviewed some medical records to further determine the nature of disagreement.

To determine overall measure agreement, we calculated the percentage of patients for whom both the claims and medical record identified at least one complication or neither identified a complication. For each case where there was a disagreement between the medical record and claims-based measure, we verified and characterized each disagreement. We then conducted a detailed review of all disagreements between the specific complications documented (or not documented) in the claims data and the medical records, even if such disagreements did not result in overall measure disagreement. We then calculated the percentage of patients where the exact complication(s) coded in claims was also documented in the medical record and vice versa (referred to throughout as “one-to-one agreement”).

Validity of Other Claims-Based Measures:

Our team has demonstrated for a number of prior measures the validity of claims-based measures for profiling hospitals by comparing either the measure results or individual data elements against medical records. CMS validated six NQF-endorsed measures currently in public reporting (acute myocardial infarction [AMI], heart failure, and pneumonia mortality and readmission measures) with models that used chart-abstracted data for risk adjustment. Specifically, claims model validation was conducted by building comparable models using abstracted medical record data for risk adjustment for heart failure patients (National Heart Failure data) (Krumholz, Wang, et al. 2006; Keenan et al. 2008), AMI patients (Cooperative Cardiovascular Project data) (Krumholz, Wang, et al. 2006), and pneumonia patients (National Pneumonia Project dataset) (Bratzler et al. 2011). When both models were applied to the same patient population, the hospital risk-standardized rates estimated using the claims-based risk-adjustment models had a high level of agreement with the results based on the medical record model, thus supporting the use of the claims-based models for public reporting.

We have also completed two national, multi-site validation efforts for two procedure-based complications measures (elective primary THA/TKA and implantable cardioverter defibrillator). Both projects demonstrated strong agreement between complications coded in claims and abstracted medical record data.

References:

Bratzler DW, Normand SL, Wang Y, et al. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. PLoS One 2011;6(4):e17401.

Keenan PS, Normand SL, Lin Z, et al. An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. Circulation 2008;1(1):29-37.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for Statistical Models Used for Public Reporting of Health Outcomes: An American Heart Association Scientific Statement From the Quality of Care and Outcomes Research Interdisciplinary Writing Group: Cosponsored by the Council on Epidemiology and Prevention and the Stroke Council Endorsed by the American College of Cardiology Foundation. Circulation. January 24, 2006 2006;113(3):456-462.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. Circulation 2006;113(13):1683-92.

Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. Circulation 2006;113:1693-1701.

National Quality Forum. National voluntary consensus standards for patient outcomes, first report for phases 1 and 2: A consensus report http://www.qualityforum.org/projects/Patient_Outcome_Measures_Phases1-2.aspx. Accessed August 19, 2010.

Shahian DM, He X, O'Brien S, et al. Development of a Clinical Registry-Based 30-Day Readmission Measure for Coronary Artery Bypass Grafting Surgery. Circulation 2014; DOI: 0.1161/CIRCULATIONAHA.113.007541. Published online before print June 10, 2014

Validity Indicated by Established Measure Development Guidelines:

We developed this measure in consultation with national guidelines for publicly reported outcomes measures, with outside experts, and with the public. The measure is consistent with the technical approach to outcomes measurement set forth in NQF guidance for outcomes measures (National Quality Forum, 2010), CMS Measure Management System (MMS) guidance, and the guidance articulated in the American Heart Association scientific statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes" (Krumholz, Brindis, et al. 2006).

Validity as Assessed by External Groups:

Throughout measure development, we obtained expert and stakeholder input via three mechanisms: regular discussions with an advisory working group, a national TEP, and a 15-day public comment period in order to increase transparency and to gain broader input into the measure.

We assembled the working group and held regular meetings throughout the development phase. The working group was tailored for development of this measure and consisted of clinicians and other professionals with expertise in biostatistics, measure methodology, and quality improvement. Working group meetings addressed key issues related to measure development, including weighing the pros and cons of and finalizing key decisions (e.g., defining the measure cohort and outcome) to ensure the measure is meaningful, useful, and well-designed. The working group provided a forum for focused expert review and discussion of technical issues during measure development prior to consideration by the broader TEP.

In addition to the working group, and in alignment with the CMS MMS, we convened a TEP to provide input and feedback during measure development from a group of recognized experts in relevant fields. To convene the TEP, we released a public call for nominations and selected individuals to represent a range of perspectives, including physicians, consumers, and purchasers, as well as individuals with experience in quality improvement, performance measurement, and health care disparities. We held three structured TEP conference calls consisting of presentation of key issues, our proposed approach, and relevant data, followed by open discussion among TEP members.

Following completion of the preliminary model, we solicited public comment on the measure through the CMS website. The public comments were then posted publicly for 30 days. The resulting input was taken into consideration during the final stages of measure development and contributed to minor modifications to the measure.

Finally, NQF previously endorsed this measure in 2012, demonstrating additional external groups' endorsement of the measure's validity.

Face Validity as Determined by TEP:

One means of confirming the validity of this measure was face validity assessed by our TEP.

List of TEP Members

1. Mark L. Francis, MD

Professor of Medicine and Biomedical Sciences, Chief, Division of Rheumatology, Department of Internal Medicine, Texas Tech University Health Sciences Center

2. Cynthia Jacelon, PhD, RN, CRRN

Associate Professor, School of Nursing, University of Massachusetts; Association of Rehabilitation Nurses

3. Norman Johanson, MD

Chairman, Orthopedic Surgery, Drexel University College of Medicine

4. C. Kent Kwoh, MD

Professor of Medicine, Associate Chief and Director of Clinical Research, Division of Rheumatology and Clinical Immunology University of Pittsburgh

5. Courtland G. Lewis, MD

American Association of Orthopaedic Surgeons

6. Jay Lieberman, MD

Professor and Chairman, Department of Orthopedic Surgery, University of Connecticut Health Center; Director, New England Musculoskeletal Institute

7. Peter Lindenauer, MD, M.Sc.

Hospitalist and Health Services Researcher, Baystate Medical Center; Professor of Medicine, Tufts University

8. Russell Robbins, MD, MBA

Principal, Mercer's Total Health Management

9. Barbara Schaffer

THA Patient

10. Nelson SooHoo, MD, MPH

Professor, University of California at Los Angeles

11. Steven H. Stern, MD

Vice President, Cardiology & Orthopedics/ Neuroscience, United Healthcare

12. Richard E. White, Jr., MD

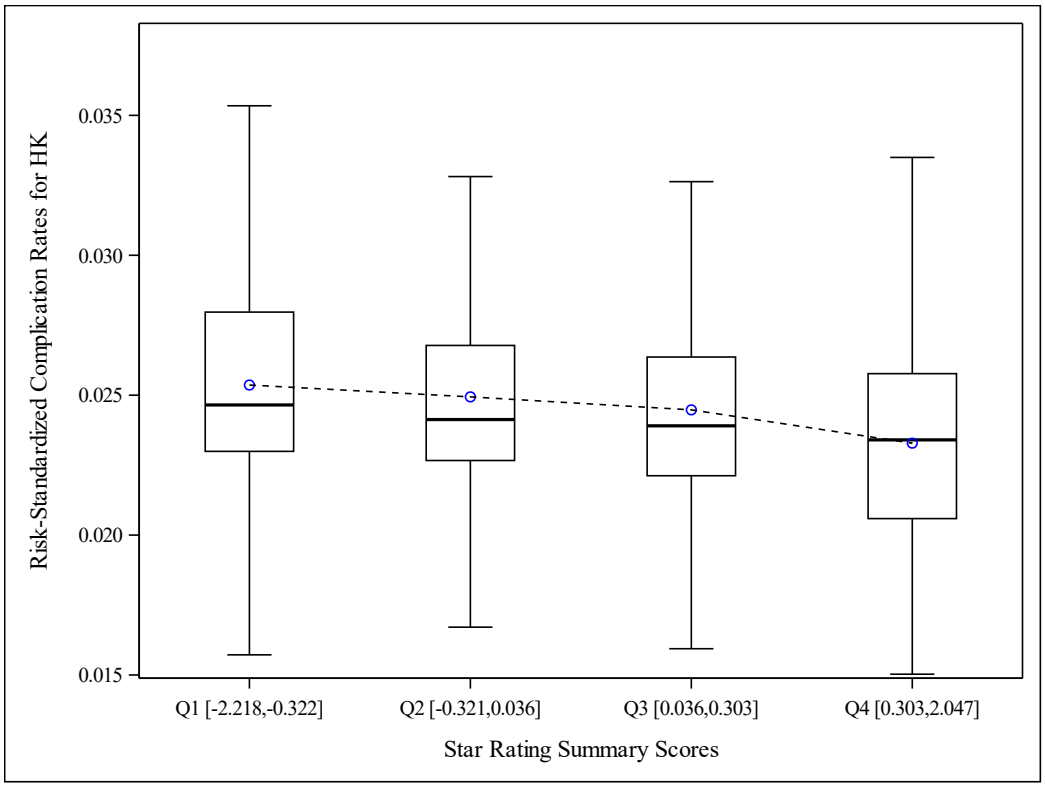
American Association of Hip and Knee Surgeons

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

Comparison to Star-Rating Summary Scores

Figure 1 shows the Box-whisker plots of the THA/TKA complications measure RSCRs within each quartile of Star-Rating summary scores. The blue circles represent the mean RSCRs of Star-Rating summary score quartiles. The correlation between THA/TKA complications and Star-Rating summary score is -0.185, which suggests that hospitals with lower THA/TKA RSCRs are more likely to have higher Star-Rating summary scores especially at the extremes.

Figure 1



Comparison to Hospital Surgical Admission Volume

Table 3 illustrates the relationship between deciles of admission volume and THA/TKA RSCRs. There is a general trend that high volume hospitals (those in the upper deciles) have lower RSCRs than hospitals in other volume deciles.

Table 3. Relationship Between Admission Volume and THA/TKA RSCRs

Measures: 25≤N (N=2,763)

Correlation coefficient between admission volumes and RSCRs: -0.25658 <.0001

Deciles of volume	# of Hospitals	Volume Range	Mean RSCR
0%~10%	283	25-43	2.51
10%~20%	274	44-67	2.54
20%~30%	273	68-100	2.54
30%~40%	276	101-144	2.62
40%~50%	276	145-200	2.55
50%~60%	274	201-278	2.55
60%~70%	275	279-380	2.49
70%~80%	280	382-528	2.36
80%~90%	276	529-804	2.35
90%~100%	276	808-9,018	2.10

Validation of Claims-Based Definition of Complications

Overall measure agreement was 93% (598/644 patients). More specifically, there were 598 patients who either had a complication coded in the claims and a complication was also documented in the medical record or who had no complication documented in both claims and medical record data. When we examined overall agreement in patients with and without complications, initial agreement was 86% for patients with a complication compared with 99% for patients without a complication. We proposed some minor changes to the measure on the basis of this validation study. Specifically, we determined that ICD-9 code 998.59, "Other postoperative infection," was not sufficiently specific to sepsis, and the measure identified cases of sepsis that were not documented in the medical record. Therefore, we recommended removal of this code from the measure specifications. Secondly, we recommended combining wound infection and periprosthetic joint infection as a single complication in the measure specifications because these complications can be clinically difficult to differentiate. After the proposed measure changes were implemented, measure agreement between claims data and the medical record will increase to 99% (635/644 patients).

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 Validity Testing

This validation approach compares the THA/TKA complication measure results against the overall star rating scores. The Figure 1 Box Plot results demonstrate an observed trend of lower risk-standardized complications

with higher star ratings, especially at the extremes, which supports measure score validity. Additionally, this validation approach compared various categories and deciles of hospital THA/TKA admission volume with THA/TKA complication measure scores in Table 3 – these results demonstrate an observed trend of higher hospital volume with lower complication measure scores. Overall, the results above show that the trend and direction of this association is in line with what would be expected.

Validation of Claims-Based Definition of Complications

The administrative claims-based and medical record data showed a high level of agreement in how they identified complications in the validity testing that was performed. There was overall measure agreement between the claims data and the medical record on the measure outcome in 99% of the cases after improving the claims-based definition of complication.

2b2. EXCLUSIONS ANALYSIS

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

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

All exclusions were determined by careful clinical review and have been made based on clinically relevant decisions to ensure accurate calculation of the measure. To ascertain impact of exclusions on the cohort, we examined overall frequencies and proportions of the total cohort excluded for each exclusion criterion (**Testing Dataset**). These exclusions are consistent with similar NQF-endorsed outcome measures. Rationales for the exclusions are detailed in data field S.9 (Denominator Exclusions).

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*)

In the **Testing Dataset (Table 4),** below is the distribution of exclusions among hospitals with 25 or more admissions:

Table 4. Frequency and Distribution of Exclusions Across Hospitals

Exclusion	N	%	Distribution across hospitals (N=2,789): Minimum, 25th percentile, 50th percentile, 75th percentile, maximum
1. Discharged against medical advice (AMA)	156	0.02	(0.0, 0.0, 0.0, 0.0, 2.94)
2. Without at least 90 days post-discharge enrollment in FFS Medicare for index admissions	7,815	0.77	(0.0, 0.0, 0.60, 1.20, 8.57)
3. Had more than two THA/TKA procedure codes during the index hospitalization	1	0.00	(0.0, 0.0, 0.0, 0.0, 0.09)

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*)

Exclusion 1 (patients who are discharged AMA) accounts for 0.02% of all index admissions excluded from the initial index cohort. This exclusion is needed for acceptability of the measure to hospitals, who do not have the

opportunity to adequately deliver full care. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 2 (patients without at least 90 days of post-discharge enrollment in FFS Medicare for index admissions) accounts for 0.77% of all index admissions excluded from the initial cohort. This exclusion is needed because the 90-day complication outcome cannot be assessed in this group since claims data are used to determine whether a patient has experienced complications. Because a very small percent of patients are excluded, this exclusion is unlikely to affect measure score.

Exclusion 3 (patients with more than two THA/TKA procedure codes during the index hospitalization) accounts for only 1 of all index procedures excluded from the initial index cohort. Although clinically possible, it is highly unlikely that patients would receive more than two elective THA/TKA procedures in one hospitalization, which may reflect a coding error.

2b3. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES
If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

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

☐ **No risk adjustment or stratification**

☒ **Statistical risk model with 33 risk factors**

☐ **Stratification by risk categories**

☐ **Other,**

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

See risk model specifications in Section 2b3.4a and the attached data dictionary.

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. This measure is risk adjusted.

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?
Selecting Risk Variables

Our goal in selecting risk factors for adjustment was to develop parsimonious models that included clinically relevant variables strongly associated with the risk of complication in the 90 days following an index procedure. We used a two stage approach, first identifying the comorbidity or clinical status risk factors that were most important in predicting the outcome, then considering the potential addition of social risk factors.

The original measure was developed with ICD-9. When ICD-10 became effective in 2015, we transitioned the measure to use ICD-10 codes as well. ICD-10 codes were identified using 2015 GEM mapping software. We then enlisted the help of clinicians with expertise in relevant areas to select and evaluate which ICD-10 codes map to the ICD-9 codes used to define this measure during development. A code set is attached in field S.2b. (Data Dictionary).

For risk model development, we started with Condition Categories (CCs) which are part of CMS's Hierarchical Condition Categories (HCCs). The current HCC system groups the 70,000+ ICD-10-CM and 17,000+ ICD-9-CM

codes into larger clinically coherent groups (201 CCs) that are used in models to predict mortality or other outcomes (Pope et al. 2001; 2011). The HCC system groups ICD- codes into larger groups that are used in models to predict medical care utilization, mortality, or other related measures.

To select candidate variables, a team of clinicians reviewed all CCs and excluded those that were not relevant to the Medicare population or that were not clinically relevant to the complication outcome (for example, attention deficit disorder, female infertility). All potentially clinically relevant CCs were included as candidate variables and, consistent with CMS's other claims-based measures, some of those CCs were then combined into clinically coherent CC groupings.

To inform final variable selection, a modified approach to stepwise logistic regression was performed. The Development Sample was used to create 1,000 "bootstrap" samples. For each sample, we ran a logistic stepwise regression that included the candidate variables. The results (not shown in this report) were summarized to show the percentage of times that each of the candidate variables was significantly associated with mortality ($p<0.01$) in each of the 1,000 repeated samples (for example, 90 percent would mean that the candidate variable was selected as significant at $p<0.01$ in 90 percent of the times). We also assessed the direction and magnitude of the regression coefficients.

The clinical team reviewed these results and decided to retain risk adjustment variables above a predetermined cutoff, because they demonstrated a strong and stable association with risk of complication and were clinically relevant. Additionally, specific variables with particular clinical relevance to the risk of complications were forced into the model (regardless of percent selection) to ensure appropriate risk adjustment for THA/TKA. These included variables representing markers for end of life/frailty, such as:

Markers for end of life/frailty:

- Decubitus Ulcer or Chronic Skin Ulcer (CC 157-CC 161)
- Metastatic and Other Major Cancers (CC 8-CC 12)
- Osteoporosis and Other Bone/Cartilage Disorders (CC 43)
- Chronic Kidney Disease, Stage 5 (CC 136)
- Hemiplegia, Paraplegia, Paralysis, Functional disability (CC 70-CC 74, CC 103, CC 104, CC 189-CC 190)
- Stroke (CC 99-CC 100)

This resulted in a final risk-adjustment model that included 33 variables.

Social Risk Factors

We weigh SRF adjustment using a comprehensive approach that evaluates the following:

- Well-supported conceptual model for influence of SRFs on measure outcome (detailed below);
- Feasibility of testing meaningful SRFs in available data (section 1.8); and
- Empiric testing of SRFs (section 2b3.4b).

Below, we summarize the findings of the literature review and conceptual pathways by which social risk factors may influence risk of the outcome, as well as the statistical methods for SRF empiric testing. Our conceptualization of the pathways by which patients' social risk factors affect the outcome is informed by the literature cited below and IMPACT Act–funded work by the National Academy of Science, Engineering and Medicine (NASEM) and the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (ASPE).

Causal Pathways for Social Risk Variable Selection

Although some recent literature evaluates the relationship between patient SRFs and the complication outcome, few studies directly address causal pathways or examine the role of the hospital in these pathways (see, for example Gopaldas et al 2009; Kim et al., 2007; LaPar et al., 2010; 2012; Trivedi et al., 2014; Buntin et al., 2017; Borza et al., 2019). Moreover, the current literature examines a wide range of conditions and risk

variables with no clear consensus on which risk factors demonstrate the strongest relationship with complication.

The social risk factors that have been examined in the literature can be categorized into three domains: (1) patient-level variables, (2) neighborhood/community-level variables, and (3) hospital-level variables.

Patient-level variables describe characteristics of individual patients, and include the patient's income or education level. Neighborhood/community-level variables use information from sources such as the American Community Survey as either a proxy for individual patient-level data or to measure environmental factors. Studies using these variables use one dimensional measures such as median household income or composite measures such as the AHRQ-validated SES index score (Blum et al., 2014; Courtney et al., 2016; Martsolf et al., 2016; White et al., 2018). Some of these variables may include the local availability of clinical providers (Herrin et al., 2015; Herrin et al., 2016). Hospital-level variables measure attributes of the hospital which may be related to patient risk. Examples of hospital-level variables used in studies are ZIP code characteristics aggregated to the hospital level or the proportion of Medicaid patients served in the hospital (Gilman et al., 2014; Joynt et al., 2013; Jha et al., 2013; Xu et al., 2018).

The conceptual relationship, or potential causal pathways by which these possible social risk factors influence the risk of complication following an acute illness or major surgery, like the factors themselves, are varied and complex. There are at least four potential pathways that are important to consider:

1. **Patients with social risk factors may have worse health at the time of hospital admission**. Patients who have lower income/education/literacy or unstable housing may have a worse general health status and may present for their hospitalization or procedure with a greater severity of underlying illness. These social risk factors, which are characterized by patient-level or neighborhood/community-level (as proxy for patient-level) variables, may contribute to worse health status at admission due to competing priorities (restrictions based on job), lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment.
2. **Patients with social risk factors often receive care at lower quality hospitals**. Patients of lower income, lower education, or unstable housing have inequitable access to high quality facilities, in part, because such facilities are less likely to be found in geographic areas with large populations of poor patients. Thus, patients with low income are more likely to be seen in lower quality hospitals, which can explain increased risk of complications following hospitalization.
3. **Patients with social risk factors may receive differential care within a hospital**. The third major pathway by which social risk factors may contribute to complications risk is that patients may not receive equivalent care within a facility. For example, patients with social risk factors such as lower education may require differentiated care (e.g. provision of lower literacy information – that they do not receive).
4. **Patients with social risk factors may experience worse health outcomes beyond the control of the health care system.** Some social risk factors, such as income or wealth, may affect the likelihood of complications without directly affecting health status at admission or the quality of care received during the hospital stay. For instance, while a hospital may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing financial priorities which don't allow for adequate recuperation or access to needed treatments, or a lack of access to care outside of the hospital.

Although we analytically aim to separate these pathways to the extent possible, we acknowledge that risk factors often act on multiple pathways, and as such, individual pathways can be complex to distinguish analytically. Further, some social risk factors, despite having a strong conceptual relationship with worse outcomes, may not have statistically meaningful effects on the risk model. They also have different implications on the decision to risk adjust or not.

Based on this model and the considerations outlined in section 1.8 - namely, that the AHRQ SES index and dual eligibility variables aim to capture the SRFs that are likely to influence these pathways (income, education, housing, and community factors) - the following social risk variables were considered for risk-adjustment:

- Dual eligible status
- AHRQ SES index

Statistical Methods

We assessed the relationship between the SRF variables with the outcome and examined the incremental effect in a multivariable model. For this measure, we also examined the extent to which the addition of any one of these variables improved model performance or changed hospital results.

One concern with including SRFs in a model is that their effect may be at either the patient or the hospital level. For example, low SES may increase the risk of complications because patients of low SES have an individual higher risk (patient-level effect) or because patients of low SES are more often admitted to hospitals with higher overall complication rates (hospital-level effect). Identifying the relative contribution of the hospital level is important in considering whether a factor should be included in risk adjustment; if an effect is primarily a hospital-level effect, adjusting for it is equivalent to adjusting for differences in hospital quality. Thus, as an additional step, we assessed whether there was a "contextual effect" at the hospital level. To do this, we performed a decomposition analysis to assess the independent effects of the SRF variables at the patient level and the hospital level. If, for example, the elevated risk of complications for patients of low SES were largely due to lower quality/higher complications risk in hospitals with more patients of low SES, then a significant hospital-level effect would be expected with little-to-no patient-level effect. However, if the increased complications risk were solely related to higher risk for patients of low SES regardless of hospital effect, then a significant patient-level effect would be expected and a significant hospital-level effect would not be expected.

Specifically, we modeled the SRF variables as follows, let X_{ij} be a binary indicator of the SRF status of the i^{th} patient at the j^{th} hospital, and $\mathbf{X}_j$ the percent of patients at hospital j with $X_{ij} = 1$. Then we added both $X_{ij} \equiv X_{patient}$ and $X_j \equiv X_{hospital}$ to the model. The first variable, $X_{patient}$, represents the effect of the risk factor at the patient level (sometimes called the "within" hospital effect), and the second variable, $X_{hospital}$, represents the effect at the hospital level (sometimes called the "between" hospital effect). By including both of these in the same model, we can assess whether these are independent effects, whether one effect dominates the other, or whether only one of these effects contributes. This analysis allows us to simultaneously estimate the independent effects of: 1) hospitals with higher or lower proportions of low SES patients on the complication rate of an average patient; and 2) a patient's SES on their own complication rates when seen at an average hospital.

It is very important to note, however, that even in the presence of a significant patient-level effect and absence of a significant hospital-level effect, the increased risk could be partly or entirely due to the quality of care patients receive in the hospital. For example, biased or differential care provided within a hospital to low-income patients as compared to high-income patients would exert its impact at the level of individual patients, and therefore be a patient-level effect.

It is also important to note that the patient-level and hospital-level coefficients cannot be quantitatively compared because the patient's SES circumstance in the model is binary whereas the hospital's proportion of low SES patients is continuous. Therefore, in order to quantitatively compare the relative size of the patient and hospital effects, we calculated a range of predicted probabilities of complications based on the fitted model.

Specifically, to estimate an average hospital effect, we calculated the predicted probabilities for the following scenarios: (1) Assuming all patients do not have the risk factor (X_{ij} =0) and hospital level risk factor is at 5% percentile (P5) of all hospital values; (2) Assuming all patients do not have the risk factor and hospital level risk factor is at 95% percentile (P95); (3) Assuming all patients do have the risk factor (X_{ij} =1) and hospital level risk factor is at 5% percentile (P5); (4) Assuming all patients have the risk factor and hospital level risk factor is at 95% percentile (P95). The average hospital effect is estimated by ((2)-(1) + (4)-(3))/2 (P95-P5). Then, to estimate an average patient effect, we first calculated the predicted probabilities by assuming patient-level risk factor equal to 0 or 1 at different hospital risk factor percentiles (0%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 100%). Then at each of those percentiles, we could obtain the difference of predicted probabilities

between all patients not having the risk factor and then all patients having the risk factor. We calculated the average of those differences in predicted probabilities ('delta') as the patient effect.

In summary, the difference in predicted probabilities at the 95th and 5th percentiles (P95-P5) estimates the hospital-level effect of the SRF on complications. The difference in predicted probabilities when all patients have and do not have the SES risk factor (delta) estimates the patient-level effect of the SES risk factor on complications. The hospital-level effect is greater than the patient-level effect when P95-P5 is greater than delta. We used P95 and P5 rather than the maximum (P100) and minimum (P0) to avoid outlier values.

We also performed the same analysis for several clinical covariates to contrast the relative contributions of patient- and hospital-level effects of clinical variables to the relative contributions for the SRFs.

References

Blum AB, Egorova NN, Sosunov EA, et al. Impact of socioeconomic status measures on hospital profiling in New York City. Circulation Cardiovascular quality and outcomes 2014; 7:391-7.

Borza T, Oerline MK, Skolarus TA, et al. Association Between Hospital Participation in Medicare Shared Savings Program Accountable Care Organizations and Readmission Following Major Surgery. Ann Surg. 2019;269(5):873-878. doi:10.1097/SLA.0000000000002737.

Buntin MB, Ayanian JZ. Social Risk Factors and Equity in Medicare Payment. *New England Journal of Medicine.* 2017;376(6):507-510.

Committee on Accounting for Socioeconomic Status in Medicare Payment Programs; Board on Population Health and Public Health Practice; Board on Health Care Services; Institute of Medicine; National Academies of Sciences, Engineering, and Medicine. Accounting for Social Risk Factors in Medicare Payment: Identifying Social Risk Factors. Washington (DC): National Academies Press (US); 2016 Jan 12. (https://www.ncbi.nlm.nih.gov/books/NBK338754/doi:10.17226/21858)

Courtney M, Huddleston J, Iorio R, Markel D. Socioeconomic Risk Adjustment Models for Reimbursement Are Necessary in Primary Total Joint Arthroplasty. July 2016; 32(1):1-5. https://doi.org/10.1016/j.arth.2016.06.050.

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation. Report to Congress: Social Risk Factors and Performance under Medicare's Value-based Payment Programs. December 21, 2016. (https://aspe.hhs.gov/pdf-report/report-congress-social-risk-factors-and-performance-under-medicares-value-based-purchasing-programs).

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

Gilman M, Adams EK, Hockenberry JM, et al. California safety-net hospitals likely to be penalized by ACA value, readmission, and meaningful-use programs. Health Aff (Millwood). Aug 2014; 33(8):1314-22.

Gopaldas R R, Chu D., "Predictors of surgical mortality and discharge status after coronary artery bypass grafting in patients 80 years and older." The American Journal of Surgery. 2009; 198(5): 633-638.

Herrin J, Kenward K, Joshi MS, Audet AM, Hines SJ. Assessing Community Quality of Health Care. Health Serv Res. 2016 Feb;51(1):98-116. doi: 10.1111/1475-6773.12322. Epub 2015 Jun 11. PMID: 26096649; PMCID: PMC4722214.

Herrin J, St Andre J, Kenward K, Joshi MS, Audet AM, Hines SC. Community factors and hospital readmission rates. Health Serv Res. 2015 Feb;50(1):20-39. doi: 10.1111/1475-6773.12177. Epub 2014 Apr 9. PMID: 24712374; PMCID: PMC4319869.

Jha AK, Orav EJ, Epstein AM. Low-quality, high-cost hospitals, mainly in South, care for sharply higher shares of elderly black, Hispanic, and medicaid patients. Health affairs 2011; 30:1904-11.

Joynt KE, Jha AK. Characteristics of hospitals receiving penalties under the Hospital Readmissions Reduction Program. JAMA. Jan 23 2013; 309(4):342-3.

Joynt KE, Orav EJ, Jha AK. Thirty-day readmission rates for Medicare beneficiaries by race and site of care. JAMA. 2011 Feb 16; 305(7):675-81. doi: 10.1001/jama.2011.123.

Kim C, Diez A V, Diez Roux T, Hofer P, Nallamothu B K, Bernstein S J, Rogers M, "Area socioeconomic status and mortality after coronary artery bypass graft surgery: The role of hospital volume." Clinical Investigation Outcomes, Health Policy, and Managed Care. 2007; 154(2): 385-390.

Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes. Circulation. 2006; 113: 456-462. Available at: http://circ.ahajournals.org/content/113/3/456.full.pdf+html. Accessed January 14, 2016.

LaPar D J, Bhamidipati C M, et al. "Primary Payer Status Affects Mortality for Major Surgical Operations." Annals of Surgery. 2010; 252(3): 544-551.

LaPar D J, Stukenborg G J, et al "Primary Payer Status Is Associated With Mortality and Resource Utilization for Coronary Artery Bypass Grafting." Circulation. 2012; 126:132-139.

Martsolf G, Barrett M, Weiss A, Kandrack R, Washington R, Steiner C, Mehrotra A, SooHoo N, Coffey R. Impact of Race/Ethnicity and Socioeconomic Status on Risk-Adjusted Hospital Readmission Rates Following Hip and Knee Arthroplasty, The Journal of Bone and Joint Surgery. 2016;98(16):1385-1391. https://doi.org/10.2106/JBJS.15.00884.

Normand S-LT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. 2007/05 2007:206-226.

Pope GC, Ellis RP, Ash AS, et al. Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Final Report to the Health Care Financing Administration under Contract Number 500-95-048. 2000; http://www.cms.hhs.gov/Reports/downloads/pope_2000_2.pdf. Accessed February 25, 2020.

Pope GC, Kautter J, Ingber MJ, et al. Evaluation of the CMS-HCC Risk Adjustment Model: Final Report. 2011; https://www.cms.gov/Medicare/Health-Plans/MedicareAdvtgSpecRateStats/downloads/evaluation_risk_adj_model_2011.pdf. Accessed February 25, 2020.

Singh JA, Kallan MJ, Chen Y, Parks ML, Ibrahim SA. Association of Race/Ethnicity With Hospital Discharge Disposition After Elective Total Knee Arthroplasty. *JAMA Network Open.* 2019;2(10):e1914259. doi:10.1001/jamanetworkopen.2019.14259.

Trivedi AN, Nsa W, Hausmann LR, et al. Quality and equity of care in U.S. hospitals. The New England journal of medicine 2014; 371:2298-308.

White, R.S., Sastow, D.L., Gaber-Baylis, L.K. *et al.* Readmission Rates and Diagnoses Following Total Hip Replacement in Relation to Insurance Payer Status, Race and Ethnicity, and Income Status. *J. Racial and Ethnic Health Disparities* 5, 1202–1214 (2018). https://doi.org/10.1007/s40615-018-0467-0.

Xu HF, White RS, Sastow DL, Andreae MH, Gaber-Baylis LK, Turnbull ZA. Medicaid insurance as primary payer predicts increased mortality after total hip replacement in the state inpatient databases of California, Florida and New York. *J Clin Anesth*. 2017;43:24-32. doi:10.1016/j.jclinane.2017.09.008.

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

☒ **Published literature**

☒ **Internal data analysis**

☐ **Other (please describe)**

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

The table below shows the final variables in the model in the testing dataset with associated odds ratios (OR) and 95 percent confidence intervals (CI).

Table 5. Adjusted OR and 95% CIs for the THA/TKA Hierarchical Logistic Regression Model over Different Time Periods in the **Testing Dataset**

Variable	04/2016-03/2017 OR (95%)	04/2017-03/2018 OR (95%)	04/2018-03/2019 OR (95%)	04/2016-03/2019 OR (95%)
Age minus 65 (years above 65, continuous)	1.03 (1.03-1.03)	1.03 (1.03-1.04)	1.03 (1.02-1.03)	1.03 (1.03-1.03)
Male	1.13 (1.08-1.18)	1.15 (1.10-1.21)	1.17 (1.11-1.23)	1.15 (1.12-1.18)
Index admissions with an elective THA procedure	1.27 (1.22-1.33)	1.30 (1.24-1.36)	1.26 (1.19-1.32)	1.28 (1.24-1.31)
Number of procedures (two vs. one)	1.70 (1.47-1.96)	1.69 (1.45-1.97)	1.65 (1.39-1.96)	1.69 (1.54-1.85)
Other congenital deformity of hip (joint)	1.66 (1.17-2.38)	1.89 (1.35-2.65)	1.25 (0.82-1.90)	1.60 (1.30-1.98)
Post traumatic osteoarthritis	1.10 (0.93-1.31)	1.13 (0.95-1.35)	1.06 (0.87-1.30)	1.10 (0.99-1.23)
Metastatic cancer and acute leukemia (CC 8)	1.04 (0.81-1.34)	0.93 (0.71-1.23)	0.91 (0.68-1.21)	0.96 (0.82-1.12)
Other major cancers (CC 9-12)	0.96 (0.90-1.02)	0.96 (0.90-1.03)	0.89 (0.82-0.96)	0.94 (0.90-0.98)
Respiratory/heart/digestive/urinary/other neoplasms (CC 13-15)	1.00 (0.95-1.06)	0.94 (0.89-1.00)	0.93 (0.87-0.99)	0.96 (0.93-0.99)
Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)	1.13 (1.08-1.19)	1.12 (1.07-1.18)	1.10 (1.04-1.16)	1.12 (1.08-1.15)
Protein-calorie malnutrition (CC 21)	2.51 (2.15-2.92)	1.69 (1.43-2.00)	1.78 (1.49-2.12)	1.97 (1.79-2.16)
Morbid obesity (CC 22)	1.65 (1.54-1.76)	1.60 (1.50-1.71)	1.64 (1.53-1.76)	1.63 (1.56-1.69)
Bone/joint/muscle infections/necrosis (CC 39)	1.10 (0.98-1.23)	1.21 (1.08-1.35)	1.40 (1.25-1.57)	1.22 (1.15-1.31)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 40)	1.14 (1.06-1.22)	1.18 (1.10-1.26)	1.23 (1.14-1.32)	1.18 (1.13-1.23)
Osteoarthritis of hip or knee (CC 42)	0.97 (0.85-1.10)	0.94 (0.83-1.08)	0.97 (0.83-1.12)	0.96 (0.89-1.04)
Osteoporosis and other bone/cartilage disorders (CC 43)	1.03 (0.98-1.09)	1.03 (0.98-1.09)	1.04 (0.98-1.10)	1.03 (1.00-1.07)
Dementia or other specified brain disorders (CC 51-53)	1.20 (1.10-1.31)	1.32 (1.21-1.45)	1.17 (1.06-1.30)	1.23 (1.17-1.30)
Major psychiatric disorders (CC 57-59)	1.43 (1.31-1.56)	1.41 (1.29-1.53)	1.36 (1.25-1.49)	1.39 (1.32-1.46)
Hemiplegia, paraplegia, paralysis, functional disability (CC 70-74, 103-104, 189-190)	1.22 (1.05-1.42)	1.17 (1.00-1.36)	1.31 (1.12-1.53)	1.23 (1.12-1.34)

Variable	04/2016-03/2017 OR (95%)	04/2017-03/2018 OR (95%)	04/2018-03/2019 OR (95%)	04/2016-03/2019 OR (95%)
Cardio-respiratory failure and shock (CC 84 plus ICD-10-CM codes R09.01 and R09.02, for discharges on or after October 1, 2015; CC 84 plus ICD-9-CM codes 799.01 and 799.02, for discharges prior to October 1, 2015)	1.24 (1.12-1.38)	1.24 (1.11-1.38)	1.39 (1.25-1.55)	1.28 (1.20-1.36)
Coronary atherosclerosis or angina (CC 88-89)	1.33 (1.27-1.40)	1.28 (1.22-1.35)	1.29 (1.22-1.36)	1.30 (1.26-1.34)
Stroke (CC 99-100)	1.01 (0.88-1.16)	1.19 (1.04-1.36)	1.12 (0.97-1.29)	1.10 (1.02-1.19)
Vascular or circulatory disease (CC 106-109)	1.14 (1.08-1.20)	1.11 (1.05-1.17)	1.16 (1.09-1.22)	1.13 (1.10-1.17)
Chronic obstructive pulmonary disease (COPD) (CC 111)	1.5 8(1.49-1.67)	1.63 (1.54-1.73)	1.55 (1.45-1.65)	1.58 (1.53-1.64)
Pneumonia (CC 114-116)	1.17 (1.07-1.29)	1.15 (1.04-1.27)	1.14 (1.03-1.27)	1.16 (1.09-1.22)
Pleural effusion/pneumothorax (CC 117)	0.99 (0.86-1.14)	0.98 (0.85-1.13)	1.00 (0.86-1.17)	0.99 (0.91-1.08)
Dialysis status (CC 134)	1.14 (0.83-1.57)	1.91 (1.45-2.53)	1.38 (1.00-1.91)	1.46 (1.22-1.74)
Renal failure (CC 135-140)	1.33 (1.26-1.41)	1.35 (1.28-1.43)	1.27 (1.20-1.35)	1.31 (1.27-1.36)
Decubitus ulcer or chronic skin ulcer (CC 157-161)	1.27 (1.13-1.43)	1.28 (1.14-1.44)	1.34 (1.18-1.51)	1.30 (1.21-1.39)
Trauma (CC 166-168, 170-173)	1.16 (1.06-1.27)	1.15 (1.04-1.26)	1.13 (1.02-1.25)	1.14 (1.08-1.21)
Vertebral fractures without spinal cord injury (CC 169)	1.03 (0.86-1.23)	0.95 (0.78-1.15)	1.11 (0.92-1.34)	1.03 (0.92-1.15)
Other injuries (modified) (CC 174)	1.13 (1.07-1.19)	1.10 (1.05-1.16)	1.12 (1.06-1.18)	1.12 (1.08-1.15)
Major complications of medical care and trauma (CC 176-177)	1.23 (1.13-1.34)	1.34 (1.23-1.46)	1.20 (1.09-1.32)	1.26 (1.19-1.32)

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.**

Throughout this section, we present new SRF testing results based on the current testing dataset (2020); in addition, we show prior analyses included in the 2016 endorsement maintenance forms for comparison purposes.

Variation in prevalence of the factor across measured entities in 2020 and 2016 (Table 6)

SRFs	2020 Prevalence % (IQR)	2016 Prevalence % (IQR)
Dual	3.40% (1.4-7.8%)	6.70% (3.9-11.7%)
AHRQ Low SES	11.7% (5.0-23.9%)	12.9% (6.4-24.0%)

The prevalence of social risk factors in the THA/TKA cohort varies widely across measured entities in 2020. The median percentage of dual eligible patients was 3.40% (IQR 1.4-7.8%) and the median percentage of patients with an AHRQ SES index score adjusted for cost of living at the census block group level equal to or below 42.7 (lowest quartile) was 11.7% (IQR 5.0-23.9%) in 2020. These results are consistent with the 2016 results presented above. Overall, there has been a decline in dually eligible and AHRQ Low SES patients since last endorsement maintenance.

Comparison of observed complication rates in patients with and without social risk in 2020 and 2016 (Table 7)

SRFs	2020 Observed Rate	2016 Observed Rate
Dual (vs. Non-Dual)	3.7% (vs. 2.4%)	4.3% (vs. 3.1%)
AHRQ Low SES (vs. SES score above 42.7)	2.9% (vs. 2.4%)	3.5% (vs. 3.1%)

The patient-level observed THA/TKA complication rates are higher for dual-eligible patients (3.7%) compared with 2.4% for non-dual patients in 2020. Similarly, the complication rate for patients with an AHRQ SES index score equal to or below 42.7 was 2.9% compared with 2.4% for patients with an AHRQ SES index score above 42.7 in 2020. For both SRF variables, patient-level complication rates have declined among all characteristic groups of patients.

Incremental effect of SRF variables in a multivariable model in 2020 and 2016

We examined the strength and significance of the SRF variables in the context of a multivariable model. When we include these variables in a multivariable model that includes all of the claims-based clinical variables, the effect size of each of these variables is moderate. In 2020, dual eligibility and the AHRQ SES index have effect sizes (odds ratios) of 1.26 and 1.16 when added separately to the model, with a slightly larger effect size than the 2016 findings (1.21 and 1.07, respectively). Furthermore, the effect size of each variable is slightly attenuated (1.24 and 1.13 for dual and SES) when both are added to the model simultaneously.

We also find that the c-statistic is essentially unchanged with the addition of any of these variables into the model (Table 8).

Table 8

THA/TKA Complications Models	2020 C-Statistic	2016 C-Statistic
Base Model: risk-adjusted model using the original clinical risk variables selected for the 2020 CMS public report of the THA/TKA complications measure	0.65	0.65
Base Model plus AHRQ Low SES based on beneficiary residential 9-digit ZIP codes (SES9) as a social risk variable	0.65	0.65
Base Model plus dual eligibility (dual) as a social risk variable	0.65	0.65
Base Model plus SES9 and dual as social risk variables	0.65	-

We find that the addition of any of these variables into the model has little to no effect on hospital performance. We examined the change in hospitals' RSCRs with the addition of any of these variables. The median absolute change in hospitals' RSCRs when adding a dual eligibility indicator is 0.005% (interquartile range [IQR] -0.004% – 0.007%) with a correlation coefficient between RSCRs for each hospital with and without dual eligibility close to 1.000. The median absolute change in hospitals' RSCRs when adding a low AHRQ SES Index score indicator to the model is 0.031% (IQR -0.006% – 0.041%) with a correlation coefficient between RSCRs for each hospital with and without an indicator for a low AHRQ SES Index score adjusted for cost of living at the census block group level is 0.982.

Contextual Effect Analysis

As described in 2b3.3a, we performed a decomposition analysis in 2020 and 2016 for each SRF variable to assess whether there was a corresponding contextual effect. In order to better interpret the magnitude of results, we performed the same analysis for selected clinical risk factors. The results are described in the tables/figures below.

Both the patient-level and hospital-level dual eligibility, and low AHRQ SES Index effects were significantly associated with THA/TKA complications in the decomposition analysis. **That the hospital level effects were significant indicates that if the dual eligible or low AHRQ SES Index variables were used in the model to adjust for patient-level differences, then some of the differences between hospitals would also be adjusted for, potentially obscuring a signal of hospital quality**.

To assess the relative contributions of the patient- and hospital-level effects, we calculated a range of predicted probabilities of complications for the SRF variables and clinical covariates (comorbidities), as described in section 2b3.3a. The results are presented in the figures and table below (table of predicted probabilities for SRF variables).

For the AHRQ SES index, the hospital-level effect (P95-P5) is greater than the patient-level effect (delta) (Figures 2 and 3; predicted probabilities for SRF variables); however, the patient-level effect (P95-P5) is greater than the hospital-level effect (delta) for dual-eligibility. For clinical variables, the patient-level effect (delta) is greater than the hospital-level effect (P95-P5) for renal failure and COPD (Figures 2 and 3; predicted probabilities for clinical variables). In sum, including SRF variables into the model would predominantly adjust for a hospital-level effect, which is an important signal of hospital quality.

In the context of our conceptual model, we find clear evidence supporting the first two mechanisms by which SRFs might be related to poor outcomes. First, we find that although unadjusted rates of complications are higher for patients of low SES, the addition of SRFs to the complications risk model, which already adjusts for clinical factors, makes very little difference. In particular, there is little-to-no change in model performance or hospital results with the addition of SRFs. This suggests that the model already largely accounts for the differences in clinical risk factors (degree of illness and comorbidities) among patients of varied SES.

Second, the predominance of the hospital-level effect of SRF variables in the decomposition analyses for 2020 and 2016 (Figures 2 and 3 below) suggests the risk associated with low SES is in large part due to lower quality of care at hospitals where more patients with these risk factors are treated; hospitals caring for socially- and economically-disadvantaged patients have higher complications risk for all of their patients. Patients with SRFs tend to receive care more frequently at lower quality hospitals compared with patients with high SRF indicators. Direct adjustment for patient SRFs would essentially "over adjust" the measure, that is to say, it would be adjusting for an endogenous factor, one that influences the outcome through the site of treatment (hospital), as much as through an attribute of the patient.

In comparison, we did not observe the same predominance of the hospital-level effect among the clinical covariates, reinforcing the sense that SRFs have a distinct causal pathway in their impact on complications risk.

Table 9. Parameter Estimates for Hospital-Level and Patient-Level in 2020 and 2016 from Decomposition Analysis

Parameter	2020 Estimate (standard error), p-value	2016 Estimate (standard error), p-value
Low SES census block group (AHRQ SES index linked to 9-digit ZIP – Adjusted for Cost of Living) – Patient Level	0.074 (0.021), <0.001	0.045 (0.018), 0.0133
Low SES census block group (AHRQ SES index linked to 9-digit ZIP – Adjusted for Cost of Living) – Hospital Level	0.732 (0.082), <.0001	0.358 (0.069), <0.0001
Dual-Eligible – Patient Level	0.176 (0.031), <.0001	0.163 (0.023), <0.0001
Dual-Eligible – Hospital Level	0.723 (0.119), <.0001	0.507 (0.090), <0.0001

Figure 2. Decomposition Analysis for 2020, THA/TKA Complications

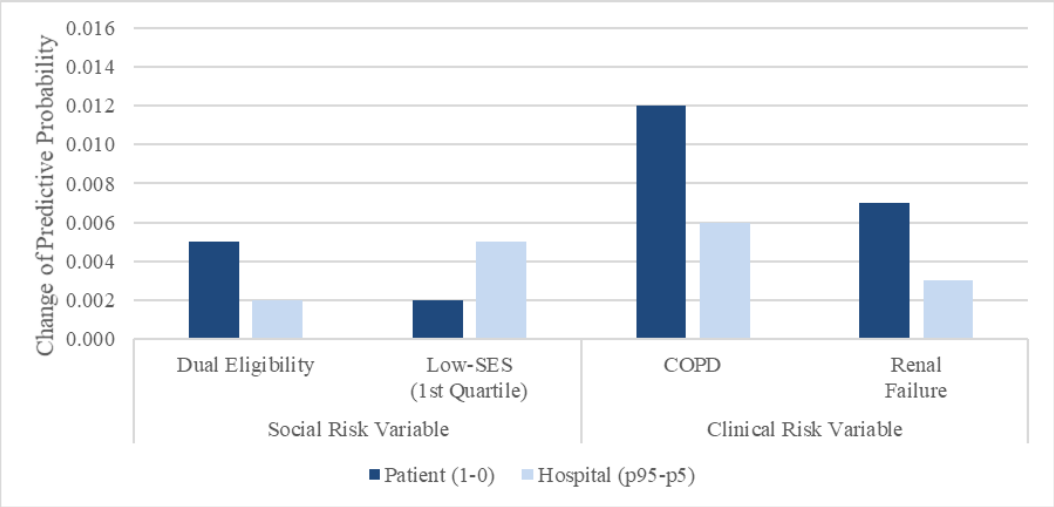
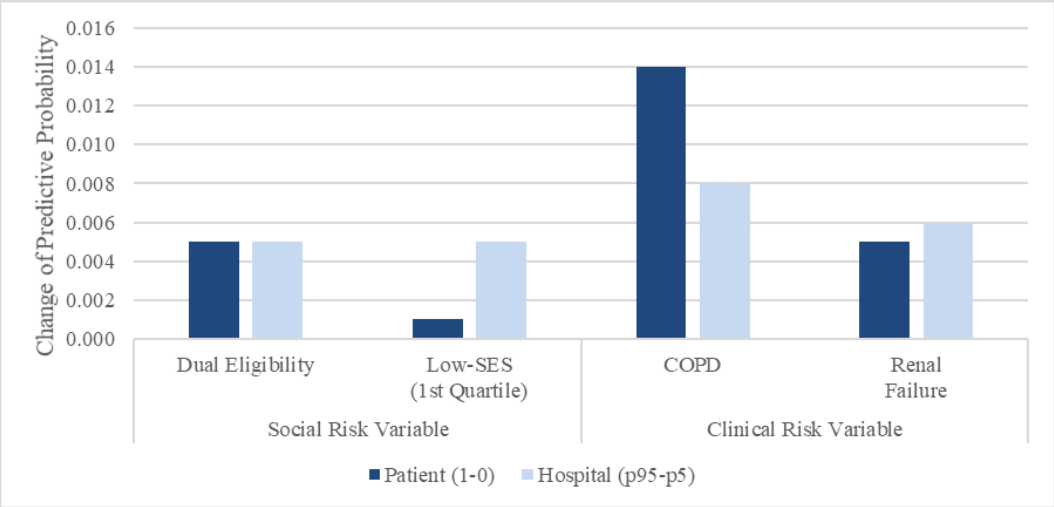


Figure 3. Decomposition Analysis for 2016, THA/TKA Complications



Summary

For risk-adjusted outcome measures, CMS first considers adjustment for clinical comorbidities, frailty indicators, and then examines additional risk imparted by SRFs after the potential for greater disease burden is included in the risk model (see section 2b3.3a). We believe that this is consistent with NQF current guidance and is appropriate given the evidence cited in our submission that people who experience greater social risk are more likely to have more disease burden compared with those who have less social risk; and that this is clearly not a signal of hospital quality. In addition, according to NQF guidance, developers should assess social risk factors for their contribution of unique variation in the outcome – that they are not redundant (NQF, 2014). Therefore, if clinical risk factors explain all or most of the patient variation in the outcome, then NQF guidance does not support adding social risk factors that account for relatively little variation. CMS's decisions about which risk factors should be included in each measure's risk-adjustment model are based on whether inclusion of such variables is likely to make the measures more successful at illuminating quality differences and motivating quality improvement. (This aim should be distinguished from decisions made in response to concerns about the impact of related payment programs on safety-net hospitals; concerns which can be addressed through other policy mechanisms.)

We found wide variation in the prevalence of the two SRFs we examined, with a large proportion of hospitals treating zero patients with these SRFs. We also found that both had some association with complication risk. However, adjustment for these factors did not have a material impact on hospital RSCRs, suggesting that existing clinical risk factors capture much of the risk related to social risk.

Ongoing research aims to identify valid patient-level social risk factors and highlight disparities related to social risk – in fact, ASPE's latest report to Congress highlights which SRFs are valid in claims data, and that adjustment for SRFs in publicly reported quality measures is not recommended because providers should be accountable for overall outcomes, regardless of social risk (ASPE 2020). As additional variables become available, they will be considered for testing and inclusion within the measure. There are alternative ways that CMS considers adjusting for social risk as part of measure program implementation, such as stratification or peer grouping. CMS also considers confidentially reporting measure disparities to hospitals so that they have more detailed, actionable information about their patient population's social risk. Given these empiric findings and program considerations, CMS chose not to include these two SRFs in the final risk model at this time.

We acknowledge the importance of balancing these competing considerations and are committed to constant refinement and improvement of risk adjustment models used in all measures. We will continue to reevaluate

this model and available risk factors on an ongoing basis, with the goal of producing the most accurate and fair risk adjustment models for assessing provider performance.

References:

Department of Health and Human Services, Office of the Assistant Secretary of Planning and Evaluation (ASPE). Second Report to Congress: Social Risk Factors and Performance in Medicare's Value-based Purchasing Programs. 2020; https://aspe.hhs.gov/system/files/pdf/263676/Social-Risk-in-Medicare%E2%80%99s-VBP-2nd-Report.pdf. Accessed July 2, 2020.

National Quality Forum (NQF). Risk adjustment for socioeconomic status or other sociodemographic factors: Technical report. 2014; http://www.qualityforum.org/Publications/2014/08/Risk_Adjustment_for_Socioeconomic_Status_or_Other_Sociodemographic_Factors.aspx. Accessed June 16, 2020.

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*)

Approach to assessing model performance

We computed three summary statistics for assessing model performance (Harrell and Shih, 2001) for the expanded cohort:

Discrimination Statistics

(1) Area under the receiver operating characteristic (ROC) curve (the c-statistic) is the probability that predicting the outcome is better than chance, which is a measure of how accurately a statistical model is able to distinguish between a patient with and without an outcome)

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; therefore, we would hope to see a wide range between the lowest decile and highest decile

Calibration Statistics

(3) Over-fitting indices (over-fitting refers to the phenomenon in which a model accurately describes the relationship between predictive variables and outcome in the development dataset but fails to provide valid predictions in new patients)

We tested the performance of the model for **the development dataset** described in section 1.7.

References:

Harrell FE and Shih YC, Using full probability models to compute probabilities of actual interest to decision makers, *Int. J. Technol. Assess. Health Care* **17** (2001), pp. 17–26.

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b3.9

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

Development and Validation Dataset

First half of randomly split development sample: C-statistic = 0.69; Predictive ability (lowest decile %, highest decile %) = (2, 15)

Second half of randomly split development sample: C-statistic = 0.70; Predictive ability (lowest decile %, highest decile %) = (2, 15)

Results for the Testing Dataset

C-statistic = 0.65

Predictive ability (lowest decile %, highest decile %): (1.1, 5.9)

For comparison of model with and without inclusion of social risk factors, see above section.

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

For the original measure development cohort, the results are summarized below:

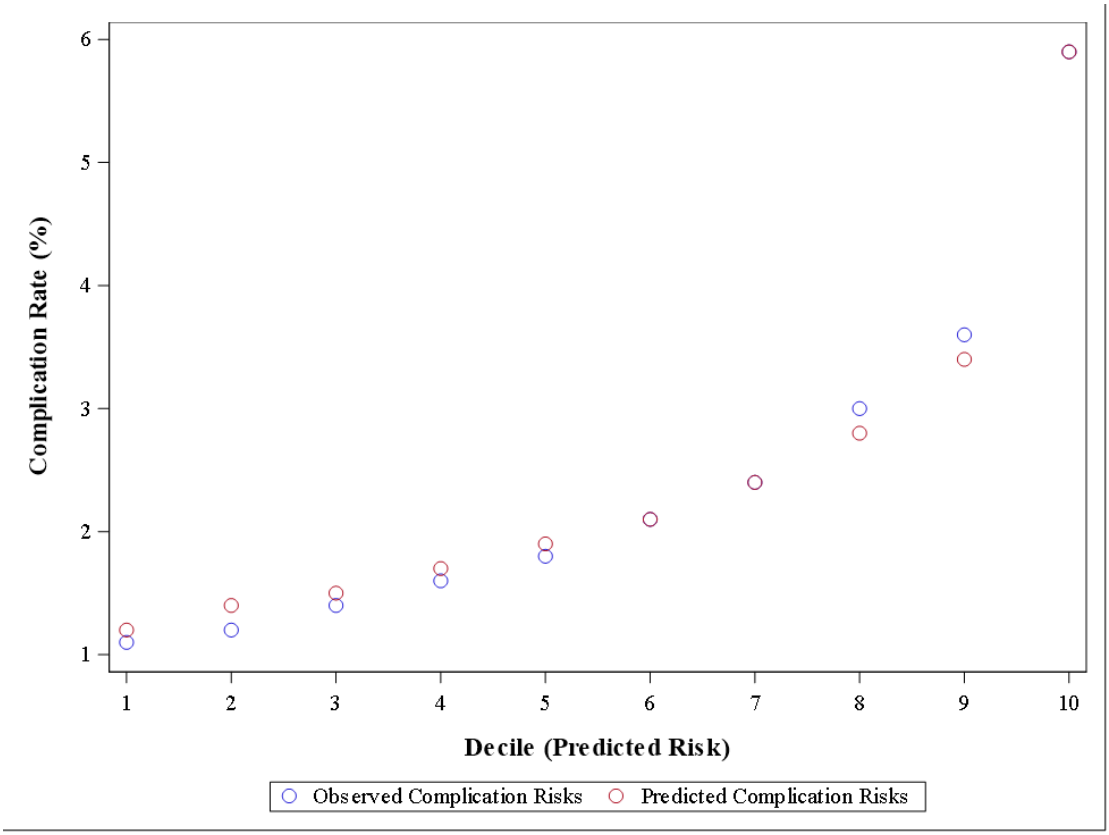
First half of split sample: Calibration: (0, 1)

Second half of split sample: Calibration: (0.04, 1.02)

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

The risk decile plot is a graphical depiction of the deciles calculated to measure predictive ability. Below, we present the risk decile plot showing the distributions for Medicare FFS data from April 2016 – March 2019 (Testing Dataset).

Figure 4. Risk Decile Plot



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*)

Discrimination Statistics

The c-statistic of 0.65 indicate fair model discrimination. The model indicated a wide range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

Calibration Statistics

Over-fitting (Calibration γ0, γ1)

If the γ0 in the validation samples are substantially far from zero and the γ1 is substantially far from one, there is potential evidence of over-fitting. The calibration value of close to 0 at one end and close to 1 to the other end indicates calibration of the model.

Risk Decile Plots

Higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. This plot indicates good discrimination of the model and good predictive ability.

Overall Interpretation

Interpreted together, our diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix) and is comparable to other outcome measures.

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

N/A

_____

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

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

The measure score is hospital-specific risk-standardized complication rates. These rates are obtained as the ratio of predicted to expected complications, multiplied by the national unadjusted rate. The "predicted" number of complications (the numerator) is calculated using the coefficients estimated by regressing the risk factors and the hospital-specific intercept on the risk of complications. The estimated hospital-specific intercept is added to the sum of the estimated regression coefficients multiplied by the patient characteristics. The results are then transformed and summed over all patients attributed to a hospital to get a predicted value. The "expected" number of complications (the denominator) is obtained in the same manner, but a common intercept using all hospitals in our sample is added in place of the hospital-specific intercept. The results are then transformed and summed over all patients in the hospital to get an expected value. To assess hospital performance for each reporting period, we re-estimated the model coefficients using the years of data in that period
We characterize the degree of variability by:

Reporting the distribution of RSCRs:

For public reporting of the measure, CMS characterizes the uncertainty associated with the RSCR by estimating the 95% interval estimate. This is similar to a 95% confidence interval but is calculated differently. If the RSCR's interval estimate does not include the national observed complication rate (because it is lower or higher than the rate), then CMS is confident that the hospital's RSCR is different from the national rate, and describes the hospital on the Hospital Compare website as "better than the U.S. national rate" or "worse than the U.S. national rate." If the interval includes the national rate, then CMS describes the hospital's RSCR as "no different than the U.S. national rate" or "the difference is uncertain." CMS does not classify performance for hospitals that have fewer than 25 cases in the three-year period.

Providing the median odds ratio (MOR) (Merlo et al, 2006). The median odds ratio represents the median increase in the odds of a complication within 30 days of a THA/TKA admission date on a single patient if the admission occurred at a higher risk hospital compared to a lower risk hospital. MOR

quantifies the between-hospital variance in terms of odds ratio, it is comparable to the fixed effects odds ratio.

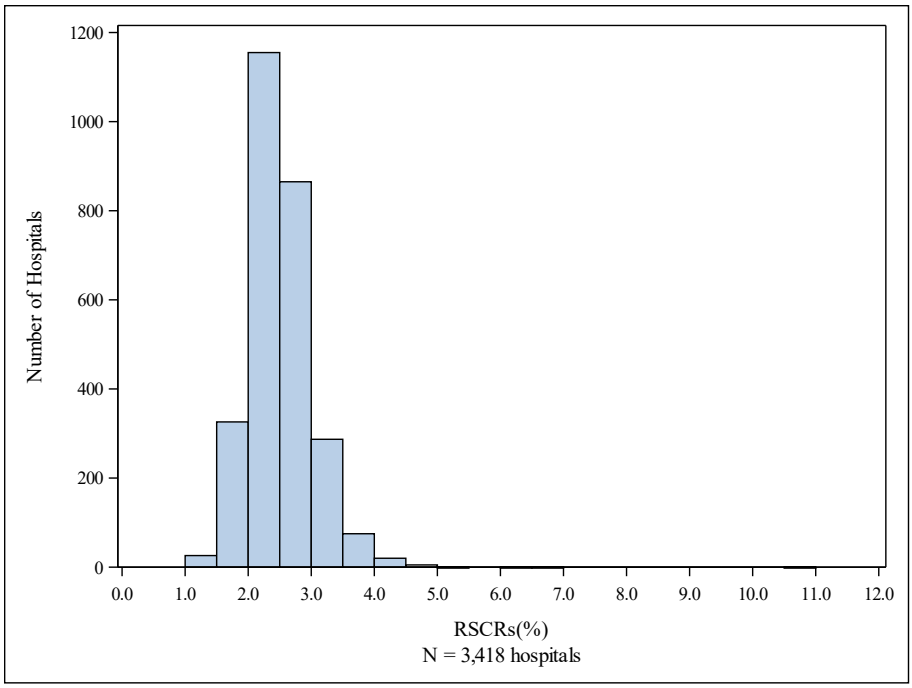
Reference:

Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Råstam L, Larsen K. (2006) A brief conceptual tutorial of multilevel analysis in social epidemiology: Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. J Epidemiol Community Health, 60(4):290-7.

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*)

Analyses of Medicare FFS data show substantial variation in RSCRs among hospitals.

Figure 5. Distribution (Histogram) Of Hospital-Level THA/TKA RSCRs



Out of 3,418 hospitals in the measure cohort, 60 performed "better than the U.S. national rate," 2,653 performed "no different from the U.S. national rate," and 50 performed "worse than the U.S. national rate." 655 were classified as "number of cases too small" (fewer than 25) to reliably tell how well the hospital is performing.

The median odds ratio was 1.38.

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 median odds ratio suggests a meaningful increase in the risk of complications if a patient has a THA/TKA procedure at a higher risk hospital compared to a lower risk hospital. A value of 1.38 indicates that a patient has a 38% increase in the odds of a complications at a higher risk performance hospital compared to a lower risk hospital, indicating the impact of quality on the outcome rate.

The variation in rates and number of performance outliers suggests there remain differences in the quality of care received across hospitals for THA/TKA procedures. This evidence supports continued measurement to reduce the variation.

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

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

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

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

N/A

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

N/A

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

N/A

2b6. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b6.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and 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 THA/TKA complications measure used claims-based data for development and testing. There was no missing data in the development and testing data.

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*)

N/A

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*)

N/A

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

Generated or collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score), Coded by someone other than person obtaining original information (e.g., DRG, ICD-9 codes on claims)

If other:

3b. Electronic Sources

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

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

ALL data elements are in defined fields in electronic claims

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

N/A

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

Attachment:

3c. Data Collection Strategy

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

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

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

This measure uses administrative claims and enrollment data and as such, offers no data collection burden to hospitals or providers.

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

N/A

4. Usability and Use

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

4a. Accountability and Transparency

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

4.1. Current and Planned Use

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

Specific Plan for Use	Current Use (for current use provide URL)
*	Public Reporting Hospital Value Based Purchasing (HVBP) Program https://www.qualitynet.org/inpatient/hvbp Hospital Value Based Purchasing (HVBP) Program https://www.qualitynet.org/inpatient/hvbp Payment Program Hospital Compare https://www.medicare.gov/hospitalcompare/search.html?

*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

Public Reporting
Program Name, Sponsor: Hospital Compare, Centers for Medicare and Medicaid Services (CMS)
Purpose: Under Hospital Compare and other CMS public reporting websites, CMS collects quality data from hospitals, with the goal of driving quality improvement through measurement and transparency by publicly displaying data to help consumers make more informed decisions about their health care. It is also intended to encourage hospitals and clinicians to improve the quality and cost of inpatient care provided to all patients. The data collected are available to consumers and providers on the Hospital Compare website at: https://www.medicare.gov/hospitalcompare/search.html. Data for selected measures are also used for paying a portion of hospitals based on the quality and efficiency of care, including the Hospital Value-Based Purchasing Program, Hospital-Acquired Condition Reduction Program, and Hospital Readmissions Reduction Program.
Payment Program
Program Name, Sponsor: Hospital Value-Based Purchasing (HVBP) Program, Centers for Medicare and Medicaid Services (CMS)
Purpose: The Hospital Value-Based Purchasing (VBP) Program is a CMS initiative that rewards acute-care hospitals with incentive payments for the quality of care they provide to people with Medicare. It was

established by the Affordable Care Act of 2010 (ACA), which added Section 1886(o) to the Social Security Act. The law requires the Secretary of the Department of Health and Human Services (HHS) to establish a value-based purchasing program for inpatient hospitals. To improve quality, the ACA builds on earlier legislation—the 2003 Medicare Prescription Drug, Improvement, and Modernization Act and the 2005 Deficit Reduction Act. These earlier laws established a way for Medicare to pay hospitals for reporting on quality measures, a necessary step in the process of paying for quality rather than quantity.
Geographic area and number and percentage of accountable entities and patients included: More than 3,000 hospitals across the country are eligible to participate in Hospital VBP. The program applies to subsection (d) hospitals located in the 50 states and the District of Columbia and acute-care hospitals in Maryland. More details about the Hospital VBP program are online at https://www.qualitynet.org/inpatient/hvbp.
The following hospitals are excluded from Hospital VBP:

- Hospitals and hospital units excluded from the Inpatient Prospective Payment System, such as psychiatric, rehabilitation, long-term care, children's, and cancer hospitals;
- Hospitals that are located in the state of Maryland participating in the Maryland All-Payer Model;
- Hospitals subject to payment reductions under the Hospital Inpatient Quality Reporting (IQR) Program;
- Hospitals cited by the Secretary of HHS for deficiencies during the performance period that pose an immediate jeopardy to patients' health or safety;
- Hospitals with an approved extraordinary circumstance exception specific to Hospital VBP; and
- Hospitals that do not meet the minimum number of cases, measures, or surveys required by Hospital VBP.

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?*)
N/A. This measure is currently publicly reported.

4a1.3. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes -- any accountability application within 3 years and publicly reported within 6 years of initial endorsement. (*Credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.*)

N/A. This measure is currently publicly reported.

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.

The exact number of measured entities (acute care hospitals) varies with each new measurement period. For the period between 2016 – 2019, all non-federal short-term acute care hospitals (including Indian Health Service hospitals) and critical access hospitals (3,418 hospitals) were included in the measure calculation. Only those hospitals with at least 25 THA/TKA procedures were included in public reporting.

Each hospital generally receives their measure results in the Spring of each calendar year through CMS's QualityNet website. The results are then publicly reported on CMS's public reporting websites in the summer of each calendar year. Since the measure is risk standardized using data from all hospitals, hospitals cannot independently calculate their score.

However, CMS provides each hospital with several resources that aid in the interpretation of their results (described in detail below). These include Hospital-Specific Reports with details about every patient from their facility that was included in the measure calculation (for example, dates of admission and discharge, discharge diagnoses, outcome [died or not], transfer status, and facility transferred from). These reports facilitate quality improvement activities such as review of individual deaths and patterns of deaths; make visible to hospitals post-discharge outcomes that they may otherwise be unaware of; and allow hospitals to look for patterns that may inform quality improvement (QI) work (e.g. among patient transferred in from particular facilities). CMS

also provides measure FAQs, webinars, and measure-specific question and answer inboxes for stakeholders to ask specific questions.

The Hospital-Specific Reports also provide hospitals with more detailed benchmarks with which to gauge their performance relative to peer hospitals and interpret their results, including comorbidity frequencies for their patients relative to other hospitals in their state and the country.

Additionally, the code used to process the claims data and calculate measure results is written in SAS (Cary, NC) and is provided each year to hospitals upon request.

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.

During the Spring of each year, hospitals have access to the following list of updated resources related to the measure which is provided directly or posted publicly for hospitals to use:

1. Hospital-Specific Reports (HSR): available for hospitals to download from QualityNet in April/May of each calendar year; includes information on the index admissions included in the measure calculation for each facility, detailed measure results, and state and national results.
2. HSR User Guide: available with the HSR and posted on QualityNet; provides instructions for interpreting the results and descriptions of each data field in the HSR.
3. Mock HSR: posted on QualityNet; provides real national results and simulated state and hospital results for stakeholders who do not receive an HSR.
4. HSR Tutorial Video: A brief animated video to help hospitals navigate their HSR and interpret the information provided.
5. Public Reporting Preview and Preview Help Guide: available for hospitals to view from QualityNet in Spring of each calendar year; includes measure results that will be publicly reported on CMS's public reporting websites.
6. Annual Updates and Specification Reports: posted in April/May of each calendar year on QualityNet with detailed measure specifications, descriptions of changes made to the measure specifications with rationale and impact analysis (when appropriate), updated risk variable frequencies and coefficients for the national cohort and updated national results for the new measurement period.
7. Frequently asked Questions (FAQs): includes general and measure-specific questions and responses, as well as infographics that explain complex components of the measure's methodology and are posted in April/May of each calendar year on QualityNet.
8. The SAS code used to calculate the measure with documentation describing what data files are used and how the SAS code works. This code and documentation are updated each year and are released upon request beginning in July of each year.
9. Measure Fact Sheets: provides a brief overview of measures, measure updates, and are posted in April/May of each calendar year on QualityNet.

During the summer of each year, the publicly-reported measure results are posted on CMS's public reporting websites, a tool to find hospitals and compare their quality of care that CMS created in collaboration with organizations representing consumers, hospitals, doctors, employers, accrediting organizations, and other federal agencies. Measure results are updated in July of each calendar year.

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.

Question and Answer Inbox (Q&A)

The measured entities (acute care hospitals) and other stakeholders or interested parties submit questions or comments about the measure through an email inbox (CMScomplicationmeasures@yale.edu). Experts on

measure specifications, calculation, or implementation, prepare responses to those inquiries and reply directly to the sender. We consider issues raised through the Q&A process about measure specifications or measure calculation in measure reevaluation.

Literature Reviews

In addition, we routinely scan the literature for scholarly articles describing research related to this measure. We summarize new information obtained through these reviews every 3 years as a part of comprehensive reevaluation as mandated by the Measure Management System (MMS) Blueprint.

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

Summary of Questions or Comments from Hospitals submitted through the Q & A process:

For the THA/TKA complications measure, we have received the following inquiries from hospitals since the last endorsement maintenance cycle:

1. Requests for detailed measure specifications including the codes used to define the measure cohort or in the risk-adjustment model;
2. Requests for the SAS code used to calculate measure results;
3. Requests about the data source used to calculate the measure;
4. Questions about how transfers are handled in the measure calculation;
5. Requests for hospital-specific measure information such as HSRs; and
6. Requests for clarification of how inclusion and exclusion criteria are applied.

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

Summary of Question and Comments from Other Stakeholders:

For the THA/TKA complications measure, we have received the following feedback from other stakeholders since the last endorsement maintenance cycle:

1. Requests for detailed measure specifications including the narrative specifications for the measure, CC-to-ICD code crosswalks, and codes used to define the measure cohort or in the risk-adjustment model;
2. Requests for the data source and the SAS code used to calculate measure results;
3. Requests for clarification of how inclusion and exclusion criteria are applied;
4. Queries about how cohorts and outcomes are defined, including how planned readmissions are defined;
5. Questions about how transfers are handled in the measure calculation; and
6. Requests for clarification on measure national rates.

Summary of Relevant Publications from the Literature Review:

Since the last endorsement cycle, we have reviewed more than 500 articles related to 90-day complications following an elective THA/TKA procedure. Relevant articles shared key themes related to: additional risk variables for consideration, including social risk factors and other clinical comorbidities; outcome rate and risk variable comparisons between inpatient and outpatient settings for both TKA and THA procedures; exploration of potential association between length of stay and THA/TKA complication outcome rates; and the relationship between complication rates and costs of care.

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.

Each year, issues raised through the Q&A or in the literature related to this measure are considered by measure and clinical experts. Any issues that warrant additional analytic work due to potential changes in the measure specifications are addressed as a part of annual measure reevaluation. If small changes are indicated after additional analytic work is complete, those changes are usually incorporated into the measure in the next

measurement period. If the changes are substantial, CMS may propose the changes through rulemaking and adopt the changes only after CMS receives public comment on the changes and finalizes those changes in the IPPS or other rule. There were no questions or issues raised by stakeholders requiring additional analysis or changes to the measure since the last endorsement maintenance cycle.

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.

The median hospital 30-day, all-cause, RSCR for the THA/TKA complications measure for the 3-year period between April 1, 2016 – March 31, 2019 was 2.4%. The median RSCR decreased by 0.1 absolute percentage points from April 2016-March 2017 (median RSCR: 2.5%) to April 2018-March 2019 (median: RSCR: 2.4%). Recent peer-reviewed literature confirms these trends (Bozic et al., 2020).

References:

Bozic K, Yu H, Zywiel MG, Li L, Lin L, Simoes JL, Sheares KD, Grady J, Bernheim S, Suter LG. Quality Measure Public Reporting Associated with Improving Hip/Knee Replacement Outcomes. JBJS 2020 (In press).

4b2. Unintended Consequences

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

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

We did not identify any unintended consequences during measure development, model testing, or re-specification. However, we are committed to monitoring this measure's use and assessing potential unintended consequences over time, such as the inappropriate shifting of care, increased patient morbidity and mortality, and other negative unintended consequences for patients.

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

N/A

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

1551 : Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

3493 : Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) for Merit-based Incentive Payment System (MIPS) Eligible Clinicians and Eligible Clinician Groups

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.

We did not include in our list of related measures any non-outcome measures (for example, process measures) with the same target population as our measure. Because this is an outcome measure, clinical coherence of the cohort takes precedence over alignment with related non-outcome measures. Furthermore, non-outcome measures are limited due to broader patient exclusions. This is because they typically only include a specific subset of patients who are eligible for that measure (for example, patients who receive a specific medication or undergo a specific procedure).

5b. Competing Measures

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

OR

Multiple measures are justified.

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

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

N/A

Appendix

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

Attachment Attachment: Procedure_Specific_Complications.pdf

Contact Information

Co.1 Measure Steward (Intellectual Property Owner): Centers for Medicare & Medicaid Services

Co.2 Point of Contact: Helen, Dollar-Maples, Helen.Dollar-Maples@cms.hhs.gov, 410-786-7214-

Co.3 Measure Developer if different from Measure Steward: Yale New Haven Health Services Corporation/Center for Outcomes Research and Evaluation (YNHHSC/ CORE)

Co.4 Point of Contact: Doris, Peter, Doris.Peter@yale.edu

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

The working group involved in the initial measure development is detailed in the original technical report available at www.qualitynet.org.

Our measure development team consisted of the following members:

Laura M. Grosso, PhD, MPH

Jeptha P. Curtis, MD

Zhenqiu Lin, PhD

Lori L. Geary, MPH

Smitha Vellanky, MSc

Carol Oladele, MPH

Yongfei Wang, MS

Elizabeth E. Drye, MD, SM

Harlan M. Krumholz, MD, SM

Working Group Members:

Daniel J. Berry, MD

Kevin J. Bozic, MD, MBA

Robert Bucholz, MD

Lisa Gale Suter, MD

Charles M. Turkelson, PhD

Lawrence Weis, MD

Technical Expert Panel Members:

Mark L. Francis, MD

Cynthia Jacelon, PhD, RN, CRRN

Norman Johanson, MD

C. Kent Kwoh, MD

Courtland G. Lewis, MD

Jay Lieberman, MD

Peter Lindenauer, MD, M.Sc.

Russell Robbins, MD, MBA

Barbara Schaffer

Nelson SooHoo, MD, MPH

Steven H. Stern, MD

Richard E. White, Jr., MD

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released: 2013

Ad.3 Month and Year of most recent revision: 03, 2019

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

Ad.5 When is the next scheduled review/update for this measure? 2020

Ad.6 Copyright statement: N/A

Ad.7 Disclaimers: N/A

Ad.8 Additional Information/Comments: N/A