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 #: 3559

Corresponding Measures:

De.2. Measure Title: Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

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

De.3. Brief Description of Measure: This patient-reported outcome-based performance measure will estimate a hospital-level, risk-standardized improvement rate (RSIR) following elective primary THA/TKA for Medicare fee-for-service (FFS) patients 65 years of age and older. Improvement will be calculated with patient-reported outcome data collected prior to and following the elective procedure. The preoperative data collection timeframe will be 90 to 0 days before surgery and the postoperative data collection timeframe will be 270 to 365 days following surgery.

1b.1. Developer Rationale: The goal of this measure is to improve patient outcomes by providing information to patients, physicians, and hospitals about hospital-level, risk-standardized patient-reported outcomes, such as pain and functional status, following elective primary THA/TKA. Measurement of patient-reported outcomes allows for a broad view of quality of care. Complex and critical aspects of care — such as communication among 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-of-care measures. As patient outcomes are not only influenced by care given during the time of hospitalization but also by patient status on presentation, outcome measures ideally are risk adjusted for patients' comorbid conditions.

THA/TKA procedures provide a particularly rich test bed for developing quality measures based upon patient-reported experiences. These procedures are commonly performed in older patients who have marked pain and functional limitation preoperatively, and who often experience significant improvements postoperatively. Patients who have undergone THA/TKA procedures have already indicated their support of such outcomes in the published literature (Liebs et al., 2013) and voiced their support for this measure as part of a TEP and a Patient Working Group.

References:

Liebs TR, Herzberg W, Gluth J, et al. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. Bone Joint J. 2013; 95-B:239–43

S.4. Numerator Statement: The numerator is the risk-standardized proportion of patients undergoing an elective primary THA or TKA who meet or exceed an a priori, patient-defined substantial clinical benefit (SCB) threshold of improvement between preoperative and postoperative assessments on joint-specific patient-

reported outcome measure (PROM) surveys. SCB improvement is defined as follows:

- For THA patients, an increase of 22 points or more on the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR); and

- For TKA patients, an increase of 20 points or more on the Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR).

SCB thresholds were defined using published literature (Lyman and Lee, 2018) and vetted by our Patient Working Group, Technical Expert Panel (TEP) and Technical Advisory Group.

References:

Lyman S and Lee YY. (2018). What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? Clin Orthop Relat Res, 467(12):2432-2441.

S.6. Denominator Statement: The cohort (target population) includes, Medicare fee-for-service (FFS) patients 65 years of age and older undergoing elective primary THA/TKA procedures, excluding patients with hip fractures, pelvic fractures and revision THAs/TKAs.

S.8. Denominator Exclusions: Patients with staged procedures, defined as more than one elective primary THA or TKA performed on the same patient during distinct hospitalizations during the measurement period, are excluded. All THA/TKA procedures for patients with staged procedures during the measurement period are removed.

De.1. Measure Type: Outcome: PRO-PM

S.17. Data Source: Claims, Instrument-Based Data

S.20. Level of Analysis: Facility

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

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

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

De.4. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results? This measure is not formally paired with another measure; however, it compliments existing outcome measures that are publicly reported on Hospital Compare including CMS' THA/TKA risk-standardized complication rate, THA/TKA risk-standardized readmission rate, and THA/TKA risk-standardized episode of care payment measures. Adding the proposed risk-standardized improvement rate in patient-reported outcomes following THA/TKA performance measure will provide a more complete picture of outcomes achieved by hospitals who perform elective primary THA and TKA procedures and will address an important measurement gap in patient-reported outcomes.

Preliminary Analysis: New Measure

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.

Criteria 1: Importance to Measure and Report

1a. Evidence

1a. Evidence. The evidence requirements for a health outcome measure include providing empirical data that demonstrate a relationship between the outcome and at least one healthcare structure, process, intervention,

or service; if these data not available, data demonstrating wide variation in performance, assuming the data are from a robust number of providers and results are not subject to systematic bias. For measures derived from patient report, evidence also should demonstrate that the target population values the measured outcome, process, or structure and finds it meaningful.

Evidence Summary

- This new, claims and instrument-based patient-reported outcome performance measure estimates a hospital-level, risk-standardized improvement rate (RSIR) following elective primary THA/TKA for Medicare fee-for-service (FFS) patients 65 years of age and older. Improvement is calculated with patient-reported outcome data collected prior to and following the elective procedure.
- Developer provided a logic model that connects the provision of high-quality THA/TKA associated care with better patient recovery, leading to decreased debilitation and better patient quality of life.
- Developer engaged with patients during the development of the measure to determine patient value and meaningfulness.
    - Patients who have undergone a THA or TKA have been engaged for input on measure development through participation on a Technical Expert Panel (TEP) and through a Patient Working Group assembled with assistance from the National Partnership for Women and Families in 2018.
    - Feedback from patients on both the TEP and the Patient Working Group indicate strong support for a patient-reported outcome-based performance measure following primary elective THA and TKA.
        - Patients stated that they expect a significant amount of improvement in both pain level and functional status following a THA/TKA procedure and felt this was an extremely important aspect of care to be captured in this measure.
        - Patients also noted that their surgical experience positively impacted not only their physical health, but their quality of life as well.
    - Developer notes that there are many studies indicating how providers can improve outcomes of the patients by addressing aspects of pre-, peri-, and postoperative care (Brown et al., 2012; Choong et al., 2009; Galea et al., 2008; Kim, 2019; McGregor et al., 2004; Moffet et al., 2004; Monticone et al., 2013; Walters, 2016).
    - Optimal clinical outcomes may be influenced by:
        - The surgeon performing the procedure
        - Team's efforts in the care of the patient
        - Care coordination across provider groups and specialties
        - Patients' engagement in their own recovery (Feng et al, 2018; Saufl et al, 2007).
    - The developer adds that the goal of hospital-level outcome measurement is to capture the full spectrum of care to incentivize collaboration and shared responsibility for improving patients' health and reducing the burden of their disease.

Question for the Committee:

- *Is there at least one thing that the provider can do to achieve a change in the measure results?*
- *This measure is derived from patient report. Does the target population value the measured outcome and find it meaningful?*

Guidance from the Evidence Algorithm

PRO-based measure (Box 1) → Relationship between the outcome and at least one healthcare action is identified and supported by the rationale (Box 2) → PASS (From Algorithm 1, NQF Measure Evaluation Criteria Sept 2019, pg. 15)

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.

- Developer provides the mean and distribution of hospital-level risk standardized improvement rates for THA/TKAs performed between July 1, 2016 to June 30, 2017 (Hospitals with >25 THA/TKA Patients with PRO Data).
  - N (Hospitals) – 123
  - Mean – 60.16%
  - Median – 66.5%
  - IQR – 18%
  - Percentile

    100% Max// 86.84%
    99%// 84.73%
    95%// 81.92%
    90%// 78.85%
    75% (Q3)// 72.51%
    50% (Median)// 66.49%
    25% (Q1)// 54.36%
    10%// 20.94%
    5%// 13.42%
    1%// 7.70%
    0%// 6.65%

Disparities

- Developer provided disparities data for n=6,734 patients within the Development Dataset, analyzing race, dual eligibility status, and socioeconomic status (SES).
- Chi-square analyses and multivariate analyses did not reveal a statistically significant association between non-White race or SES.
- Dual eligibility was borderline significant (p=0.058) at the bivariate level and statistically significant within the risk model.

Questions for the Committee:

- Is there a gap in care that warrants a national performance measure?
- Are you aware of evidence that disparities exist in this area of healthcare other than those analyzed by the developer?

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.

- Patient feedback was utilized during the design of this measure. Evidence is provided to confirm the need for this measure.
- Measure applies directly to patient-oriented outcome of reduced pain and increased function post hip/knee replacement; comment regarding dual eligibility is statistically significant confusing- perhaps dual eligibility has a different meaning in this context than clinically used.
- Pass.
- Logic model is presented. One or more processes under the control of the hospital influences outcome.
- Extensive input about importance and meaningfulness from patients and TEP; patients emphasized importance of measure to them.
- Acceptable indirect evidence.
- My 91 year-old mother needed emergency surgery after she fell and broke her hip. While this is not an elective case situation, her experience illuminates issues with what to measure. As noted in Evidence (subcriterion 1a), "Optimal clinical outcomes depend not just on the surgeon performing the procedure, but also on: the entirety of the team's efforts in the care of the patient; care coordination across provider groups and specialties; and the patients' engagement in their recovery (Feng et al, 2018; Saufl et al, 2007). Immediately after surgery the doctor told me her bones were porous, but the surgery went great. Within a week she was screaming in pain. It took another 10 days to discover the issue. I had to take her to the hospital for additional testing. The care coordination fell in my lap. I needed care transport, an aide and to set up appointments. Each bump in the transport vehicle sent vibrations to her hip and she'd scream out in pain. As the medical professionals moved her into varying positions to take pictures, she screamed in pain exclaiming she'd rather die. Turns out the screws placed in her hip slipped 8 millimeters into her hip socket. I was told she couldn't have it fixed unless she passed a cardiac exam for a second surgery. She needed to go back the next day for more testing to approve her heart was able to withstand a second surgery. I told the surgeon that I'd sign a form stating I would not sue so she would not have to undergo the cardiac tests. It was a form of human torture to have her live with a slipped screw in her hip socket. I told him just the transporting her to the hospital caused her pain. My mother screamed as she had to maneuver rooms for the cardiac tests. The Cardiac doctor looked directly at me and said it is protocol, the surgeon does not want to be sued if she dies from a cardiac issue. As my mom cried, I looked back at him and said do you hear yourself? She passed the cardiac exam and the next morning the surgeon did a full hip replacement. The Surgery cost $90,000. I don't know what the cardiac test cost or the first surgery. It is six months later and she feels great, now. From a physical ability she is 7 years younger. This measure is for Medicare fee-for-service (FFS) patients 65 years of age and older. In my mother's case there were several costly steps that could have possibly been avoided. My mother endured two surgeries within a three-week period. If she had a total hip replacement, when her hip broke, Medicare would have only paid for one surgery. In a fee-for-service model, overtreatment needs to be part of the output measure. As stated, the developer adds that the goal of hospital-level outcome measurement is to capture the full spectrum of care to incentivize collaboration and shared responsibility for improving patients' health and reducing the burden of their disease. Therefore, the outcome needs to be measured as part of the patient experience and not just the point in time. This is referred to "experience with care." The exclusion of patients with staged procedures, defined as two or more elective primary THA or TKA procedures performed on the same patient during distinct hospitalizations during the measurement period eliminates capturing data looking at an overall experience perspective. Which is one of the goals of the measures. The developer acknowledges " 1) the recovery from one procedure may

negatively impact recovery from the other procedure; and 2) it may be challenging to fully distinguish the recovery for either of the procedures from the other with postoperative PRO data (collected 270 to 365 days after surgery). This is important because it is a way to capture opportunities for improvement. For example, it is unclear to me if they measure the porous level of bones when deciding if the hip will be fully replaced or a rod used to fix. For this measure of elective surgery, does the porous nature of bones affect the outcome? How many people have bad results because their bones cannot hold the replacement parts and end up excluded because they had repeat surgeries? Table 2 Risk Variables Collected with Patient -Reported Outcome data evaluates BMI but not bone porous levels. In Table 3 Musculoskeletal System Involvement it is ranked high, but I am not sure this is about porous nature. I am not a doctor but from my review there are many criteria for evaluating the patients, smoking, narcotic use commodities, anxiety, but porous nature of bones seems to be missing. The developer states that criteria was developed using patients, literature, doctors yet, two factors that do not seem to be addressed are bone density and device malfunction. I understand my mother is a datapoint of one. Today, my mother would give the highest scores for decrease in pain and improved mobility and quality of life, however, the interim steps like improving communication among providers involved in care transition, educating patients about the best practices to recovery and rehabilitation needed improvement and therefore I support the that a hospital-level measure is a good way to encourage communication across providers to improve coordination of care at a facility overall. Also, how do we track if the product is faulty? It is unclear to me how we can use this measure to track device issues and outcomes. It may provide insight to product problems. The target population would find this meaningful.

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?

- Performance data was provided including subgroups.
- Limited data on the measure by population subgroups, perhaps more is included in documents I am unable to access, comment on dual eligibility as above.
- High.
- Gaps in performance exist.
- 2016-2017 data - opportunity for improvement (60% mean for reaching threshold); disparities for dually eligible population.
- Evidence supports gap.
- Yes. The developer states that they will continue to assess the impact of social risk for this measure over time based on disparities. This article may provide additional insight: Racial and Ethnic Disparities in Hip and Knee Joint Replacement: A Review of Research in the Veterans Affairs Health Care System https://pubmed.ncbi.nlm.nih.gov/17766799/.

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability: Specifications and Testing

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

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

Reliability

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

2a2. Reliability testing demonstrates if the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population 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: Scientific Methods Panel Group 1

Methods Panel Review (Combined)

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel and discussed on the call. A summary of the measure and the Panel discussion is provided below:

- Reliability: H-5; M-1; L-2; I-1 (Pass)
- Validity: H-0; M-5; L-3; I-0 (Pass)

In their preliminary analyses the subgroup members found the measure to be reliable, but consensus was not reached on validity. Reviewers identified several concerns related to missing data, exclusions, and the attribution approach. One panel member also raised concern regarding the impact of this measure given the selection of outcome measures, HOOS, JR and KOOS, JR, on the measurement landscape, alignment with registries, and other similar approaches. Developers provided a detailed response to the reviewers' concerns on these issues including a summary of the development process which relied heavily on technical experts and patients in particular; this process guided the selection of the patient reported outcome measure/instrument. The developers clarified the rationale for minimum case size of 25 per hospital, and the exclusions of staged procedures. The developers also noted support for this measure among orthopedic societies. Some panel members questioned whether this measure should be considered a composite. After weighing these concerns with the developers' responses, the panel passed the measure on validity.

Reliability

- Reliability testing conducted at the data element and score level

- Data element reliability testing assessed consistency and test-retest reliability of the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR) and Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR) instruments.
    - HOOS, JR internal consistency using Person Separation Index (PSI) was 0.86 and 0.87 in the two cohorts tested.
    - HOOS, JR test-retest results produced ICCs between 0.75 and 0.97.
    - KOOS, JR internal consistency using PSI was 0.84 and 0.85 in the two cohorts tested.
    - KOOS, JR test-retest results produced ICCs between 0.75 and 0.93.
- Score level reliability testing consisted of a signal-to-noise analysis. Results from a sample of 123 hospitals yielded a mean of 0.95 and a range from 0.90 to 0.99.
- Notes and results, concerns of SMP on reliability and specifications:
    - Measure specifications: Some NQF Panel members wanted clarification on the measure result calculation and definitions for "predicted," "expected" and "overall observed" improvement.
    - Data element reliability: Concern was expressed about data element reliability testing for "critical data elements" other than the HOOS, JR and the KOOS, JR. Data elements of concern were noted to be those that "make-up the denominator:" the two additional PRO tools used in the risk model and additional risk factors, including the clinical characteristics based on coding (e.g. liver disease, severe infection).
    - Reliability impact of proxy surveys: An NQF Panel member voiced concern about proxy assessment, noting that it "is unorthodox and can add significant noise."
    - Measure conversion: An NQF Panel member noted that the HOOS, JR and KOOS, JR appeared to have been transformed from 0-100 but no specifications on the approach to transformation were provided.
    - Score change calculation: An NQF Panel member noted that the interval over which the "change" in score appears to have been estimated (90-0 days prior to surgery and 270-365 days following surgery) is quite wide and could vary for an individual patient by as much as 6 months.
    - Exclusions: There was a request for clarification about how the measure accounts for patients that die between the hospital discharge and the postoperative PRO data collection period (270-365 days postoperatively), and whether they are considered "lost to follow-up." Another NQF Panel member noted that excluding deaths seemed reasonable but suggested a check on death as a possible adverse event.

Validity

- Validity testing was conducted at both the data element and score levels
- Data element validity testing included responsiveness, external validity, floor and ceiling effects for both HOOS, JR and KOOS, JR. HOOS, JR responsiveness produced standardized response means relative to other PROMs (HOOS domains, The Western Ontario and McMaster University Arthritis Index [WOMAC] domains) measuring post-surgery hip improvement of 2.38 and 2.03 in the two samples.
    - HOOS, JR external validity used Spearman's correlation analysis with the HOOS and WOMAC instruments and produced 0.87 for both samples.
    - HOOS, JR showed floor (0.6%–1.9%) and ceiling (37%–46%) effects and were comparable to or better than HOOS domains and the WOMAC.
    - KOOS, JR responsiveness produced standardized response means relative to other PROMs (KOOS, WOMAC) measuring post-surgery hip improvement of 1.79 and 1.70 in the two samples.
    - KOOS, JR external validity used Spearman's correlation analysis with the KOOS and WOMAC instruments and produced 0.89 and 0.91 for the two samples.
    - KOOS, JR showed floor (0.4%–1.2%) and ceiling (18.8%–21.8%) effects.
- Score level validity testing included empirical comparisons to another quality measure: NQF 1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA.

Comparison of THA/TKA PRO-PM RSIRs to RSCR categories indicates an increasing monotonic trend. Those hospitals in the "RSCR Worse than National Average" category have lower median RSIRs (51.87%) than the median RSIR (66.49%) of hospitals in the "RSCR Same as National Average" category, which is lower than that of hospitals in the "RSCR Better than National Average" category (71.13%).

- Notes and results, concerns of the SMP on Validity:
    - Attribution: An NQF Panel member noted concern about attributing changes in joint function to the hospital (versus care such as rehabilitation services) with a follow-up interval of nine months to one year following surgery.
    - "Unstaged procedures": An NQF Panel member suggested that the exclusion of staged procedures might eliminate up to 43% of procedures, and that the measure name should include "from unstaged procedures."
    - Exclusion analysis: An NQF Panel member noted concern that data were not provided on how the excluded patients impacted the performance measure scores.
    - Exclusion thresholds: Some NQF Panel members had questions about the 25-case volume threshold—what the threshold was based on, what happens to a facility that falls below the 25-case recommendation, if facilities without 25 cases would be excluded from the measure (and should be identified as an exclusion), and if excluded, whether it would create an incentive for them to not complete data.
    - Risk-adjustment: "The model was developed including cases from hospitals not used for reliability, validity, and missing data testing, i.e., hospitals with low caseloads (n<25) not recommended for this measure. Did the developers do a sensitivity test to assess the impact of excluding these hospitals from the risk-adjustment development sample on the risk-adjustment model?"
    - Meaningful differences: An NQF Panel member requested clarity for the data provided and whether there are meaningful differences between hospitals in the top quartile.
    - Missing data: Two NQF Panel members voiced concerns about missing data and that the only complete data were analyzed without accounting for what is likely "fairly extensive missingness." One of these members noted concern that missing surveys were accounted for but that missing responses within the survey were not.

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 (SMP) is satisfied with the reliability testing for the measure. Are there additional items related to the reliability of the measure that should be discussed?

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

Questions for the Committee regarding composite construction:

- Do you have any concerns regarding the composite construction approach (e.g., do the component measures fit the quality construct and add value to the overall composite? Are the aggregation and weighting rules consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible?)?
- The [staff] or [Scientific Methods Panel] is satisfied with the composite construction. Does the Committee think there is a need to discuss and/or vote on the composite construction approach?

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

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.
- High.
- Moderate, case-mix adjustment not clear.
- Measure the overall experience because the developer builds the case that there are factors beyond surgery outcome that impact the experience. So, the measure as designed can be consistently implemented but I am unclear how we capture communication, coordination and faulty devices and tie it back to the overall patient journey. Improvement steps may need to be taken, not just as a result of the outcome of the surgery but as the process of care.
- See comments from Scientific Methods Panel, questions addressed re missing data, exclusions, timeframe, attribution.
- No concerns.
- No significant concerns.

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

- No.
- As designed, the reliability of the measure is valid. The issue is with exclusions.
- No concerns.
- No.
- Methods panel passed, but mixed. Given measure complexity and mixed review, although "passed" consider return to methods panel with explanations for review.
- Questions addressed in Scientific Methods Panel review - test-retest, internal consistency reliability within acceptable ranges.
- No.

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

- No significant concerns.
- No concerns.
- Moderate.
- Validity testing needs clarification in meeting presentation.
- As designed, the testing results are valid. There are issues with missing data.
- See Scientific Methods Panel Review - extensive questions about validity- appear to be satisfactorily addressed.
- Agree-moderate.

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?

- No concerns.
- Obviously, patients who die or are lost to follow up may represent the population who has had the worst outcome as a result of their surgery, but this population will be mixed and should be tracked elsewhere.
- None.
- Again, given mixed review by the methods panel for this complex measure, consider return developer response to methods panel for review. The wide possible variation in time across respondents in pre-op and post op raises concerns. Also please explain rationale for such a long post-op window given trade-offs with attribution and increasing risks for intervening events. Methods panel raised other important considerations that should be addressed.
- Patient groups are excluded from the measure. According to results less than 5% of patients have multiple surgeries. While this is a small number, the practice of exclusion from the data set may limit information about potential harms occurring at a facility. If these patients have experienced preventable medical harm, 5% is not a small number but a large number.
- Risk adjusted-appropriate.
- Acceptable.

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?

- Not significant.
- No concerns.
- None.
- Please clarify why a fixed change score was selected regardless of starting point. Please compare outcomes to other instruments and address the absence of walking on the abbreviated outcome measure.
- Constitute a threat - adequately addressed by Scientific Methods Panel.
- No.
- Yes, the exclusion of patients with staged procedures may impact the validity of the measure. Patients who experienced at least one of eight complications within 90 days of the procedure are not included because the overlapping recovery periods for staged procedures occurring within one year of each other has two consequences that set patients experiencing staged THA/TKA procedures apart from patients experiencing unilateral or bilateral procedures: 1) the recovery from one procedure may negatively impact recovery from the other procedure; and 2) it may be challenging to fully distinguish the recovery for either of the procedures from the other with postoperative PRO data (collected 270 to 365 days after surgery). For patients that are harmed and have to undergo an additional surgery, eliminating this data excludes the ability to determine if a repeatable and preventable medical issue is occurring. Therefore, the hypothesis is not conceptually sound. Plus, eliminating known complications within 90 days of the procedure does not give a comprehensive review. This may not be an appropriate method. As stated in Note 14, Risk factors that influence outcomes should not be specified as exclusions. In this case, the exclusions are less

than 5%, but the notes do not quantity elimination based on small numbers. In this case, the data set intentionally omits problem points.

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.

- Developer notes that most providers in the CMMI Comprehensive Care for Joint Replacement model are not using electronic documentation of patient responses for PROs; they are using paper.
- Developer states that "incentivized PRO data collection within CMS's Comprehensive Care for Joint Replacement (CJR) model presents proof of concept for feasible, low burden collection of PROs for hospital-level quality measurement."

Questions for the Committee:

- Is this PRO-PM burdensome to patients or providers?
- 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?

- Challenge of locating patients months after surgery.
- Low burden to providers or patients.
- Challenge of locating patients' months after surgery.
- Moderate.
- It is reasonable to expect providers to ascertain baseline functional ability and limitations prior to procedure and after reasonable rehabilitation period since goal is to improve. It is not clear whether these particular items/scales are commonly used in current practice.
- As stated by the developer, "the definition of burden is subjective." Based on the design and input from the developer the measurement is based on data elements are routinely generated and used during care delivery with some of the required data elements are available in electronic form. For those still using a paper collection process, transition to on-line will need to move forward. The data collection strategy is ready to be put into operational use.
- Moderate.
- Agree with measure worksheet.

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**

OR

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

Accountability program details

- In response to the question of a credible plan for implementation, developer responds: This PRO-PM will be implemented in to-be-determined federal accountability programs through rulemaking in the future.

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

- Developer notes that this measure has not been implemented and thus they do not have feedback from those being measured. However, they have tested the measure and could have potentially obtained feedback regarding implementation from those in the testing phase. This was not included in the submission

Additional Feedback:

- Developer cites feedback from the TEP discussed in the evidence section and notes largely positive responses from the TKA/THA patients who participated.

Questions for the Committee:

- How have (or can) the performance results be used to further the goal of high-quality, efficient healthcare?
- How has the measure been vetted in real-world settings by those being measured or others?

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

RATIONALE: Developer did not present a credible plan for implementation. This criterion is not must-pass for new measures.

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

- No improvement results to report on this new measure

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 None reported for this new measure

Potential harms None reported for this new measure

Additional Feedback: None reported for this new measure

Questions for the Committee:

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

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

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

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.

- No concerns.
- There has been no opportunity for use/feedback.
- The clinical criteria for elective hip or knee replacement is pain and loss of function. Surgeons may be less likely to operate on patients who they determine are less likely to have clinical improvement due to other comorbidities.
- Moderate.
- The time frames raised concern about identifying specific areas for performance improvement.
- The benefits of data at the hospital level provides information to improve the quality of care and organizational performance.
- No obvious unintended consequences.
- There has been no opportunity for use/feedback.

Criterion 5: Related and Competing Measures

Related or competing measures

- The developer provides the following list of related measures:

- 0422 : Functional status change for patients with Knee impairments
- 0423 : Functional status change for patients with Hip impairments
- 0424 : Functional status change for patients with Foot and Ankle impairments
- 0425 : Functional Status Change for Patients with Low Back Impairments
- 0426 : Functional status change for patients with Shoulder impairments
- 0427 : Functional status change for patients with elbow, wrist and hand impairments
- 0428 : Functional status change for patients with General orthopaedic impairments
- 1550 : Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
- 1551 : Hospital-level 30-day risk-standardized readmission rate (RSRR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)
- 2643 : Average change in functional status following lumbar spine fusion surgery
- 2958 : Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery

- Staff did not identify additional measures

Harmonization

- Developer notes that the measure is harmonized with the related measures to the extent possible.

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?

- Multiple related measures are noted.
- No.
- No.
- Competing measures exist. Developers argue that many are at provider level as opposed to hospital level.
- Measure is harmonized with other measures as stated by the developer.
- 11 related measures listed - functional status changes, complication rates and readmissions - relevant and appear to be harmonized as much as possible.
- No.

Public and Member Comments

Comments and Member Support/Non-Support Submitted as of: Month/Day/Year

- **Of the XXX NQF members who have submitted a support/non-support choice:**
  - XX support the measure
  - YY do not support the measure

Scientific Acceptability: Preliminary Analysis Form

Measure Number: 3559

Measure Title: Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/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**
Panel Member #9: Additional data from various sources used for risk adjustment

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: Specifications, 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: I have some concerns with the numerator specifications:

   - It is stated within different sections of the specifications that the measure is both risk-stratified and risk-adjusted. No definitions were provided for these two terms and the difference between them.
   - I could not identify a clearly defined risk-adjustment approach within the specifications, other than stating that predicted and observed improvements are a result of an HLM that adjust for patient case mix.
   - It seems that the numerator is defined by a ratio of predicted and observed improvements multiplied by observed improvement. These three terms are not clearly defined within the specifications. I am guessing that observed improvement might be the percent of patients achieving a predetermined threshold of change from pre to post-operative status that suggest substantial improvement identified using an anchored based approach developed in a 2018 publication. However, while reading the specifications, as mentioned, some guessing is involved as not all elements needed to compute the numerator are defined.

- There seems to be a caseload threshold of 25 cases recommended. Shouldn't this be a threshold for the denominator, and what is this threshold based on?
- After reading more details on risk-adjustment methods within the testing form, the numerator calculation method starts to clarify. It is important to have all information within the specifications, or clear references to specific sections of the testing form.
- I find the lengthy description of multiple considerations used to define the numerator distracting and confusing. If I understand correctly, some of these considerations were rejected during the development phase. It may be best to only describe considerations that support the proposed numerator.
- There seems to be an additional exclusion criterion that is not defined as such, which is having at least 25 cases per measurement year, as reliability, validity and missing data testing was conducted only using hospitals exceeding this threshold, and developers note that "we therefore recommend this measure be reported using a minimum case-volume cut-off of 25 or greater".

A few minor points

- S.2d has a typo?
- S.3.1 should be NA since this is not a maintenance submission

Panel Member #3: For the numerator, the measure developer recommends facilities have 25 or more cases that have complete PRO data and risk variable data but does not indicate what happens if a facility falls below the 25 case recommendation. Are those facilities excluded? And if these facilities are excluded, does that create an incentive for them to not aim for complete data? It is also unclear how the measure accounts for patients that die between the hospital discharge and the postoperative PRO data collection period (270-365 days postoperatively). My guess is they are considered to be "lost to follow-up" and therefore would not count toward the 25 case count noted above as they would have incomplete data. This could be clearer.

Panel Member #4: No concerns

Panel Member #5: None

Panel Member #7: The scoring of HOOS JR and KOOS JR surveys appears to have been transformed from 0-100; the "change" in points on both appear to reflect that transformation but no specifications on the approach to transformation were provided. The interval over which the "change" in score appears to have been estimated (90-0 days prior to surgery and 270-365 days following surgery) is quite wide and could vary for an individual patient by as much as 6 months. With an interval of 9 mos to 1 yr following surgery, attribution of changes in joint function to care provided by the hospital (vs. e.g. rehabilitation services) appears problematic. And this appears to be a composite measure, but that NQF form does not appear to have been completed.

Panel Member #8: I found S.14, creation of the numerator, to be difficult to follow. Specifically, the following sentence "Hospital-specific risk-standardized improvement rates (RSIRs) are calculated as the ratio of a hospital's "predicted" improvement to "expected" improvement multiplied by the overall observed improvement rate." I would like to see examples for how to calculate these to be assured that the measure will be calculated in a reliable way.

I was surprised to see the specifications in S5 for the risk adjustment measure as the variables listed do not seem to match the discussion in 2b3.

Panel Member #9: None noted.

RELIABILITY: TESTING

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

Panel Member #10: Assessed both (1) internal reliability and (2) test-test reliability of survey-based outcome measure. These reliability measures were high for THA/TKA. Results from this testing (i.e. test-retest reliability > 0.70) consistent with data element reliability.

Also assessed score-level reliability using SNR which was 0.96 consistent with excellent score-level reliability.

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**

Panel Member #1: The reason for marking 'NO' is that hospitals with less than 25 cases were not included, a criterion not specified in the exclusions for this measure.
Panel Member #2: Developers report that HOOS JR was not tested for reliability because HOOS was several times, and do not state it was tested here. KOOS, on the other hand, was stated as tested for test-retest (stability). Reliability data are presented for legacy HOOS and KOOS (internal consistency and stability). Measure score reliability was tested using signal-to-noise (Adams) method.

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: I have no concerns with the methods reported and used for testing reliability.
   Panel Member #2: Methods appropriate
   Panel Member #3: It is a bit confusing on which dataset was used for the reliability testing --- I believe it is the 123 hospitals in the "Combined Dataset"? It is unclear how the Development and Validation data sets are used.
   For data element reliability testing, the measure developer provides published reliability data from the HOOS, Jr. and KOOS, Jr., but did not provide any testing for a number of other critical data elements (i.e., data elements that make-up the denominator; the two additional PRO tools used in the risk model; additional risk factors).
   The signal-to-noise approach for measure score reliability is an appropriate method.
   Panel Member #4: The testing method appears adequate for data element level and measure score level.
   "Data Element Reliability
   HOOS, JR Reliability:
   Internal consistency: assessed internal consistency reliability of using the Person Separation Index (PSI). The PSI was used in two data samples, the Hospital for Special Surgery (HSS) cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR), a nationally representative joint replacement registry. A higher value on the PSI indicates greater ability to differentiate patients with varying levels of ability, which in turn provides evidence of good internal consistency. For testing internal consistency for the HOOS, JR, a PSI value greater than 0.7 was considered acceptable
   KOOS, JR Reliability:
   Internal consistency: [same as above] [p8]
   Measure Score Reliability: Using the Combined Dataset (Development and Validation Datasets), we identified the hospitals with at least 25 THA/TKA patients with PRO data during the measurement period and assessed signal-to-noise reliability…. [Z]ero 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." [p9]

Panel Member #7: Reliability testing appears to have been done at the patient not hospital level (the unit of comparison of the measure)

Panel Member #8: For data element reliability, internal consistency person separation index and test-retest reliability (ICCs) were assessed. I have no concerns with the methods used.
For Measure score reliability, a signal to noise ratio was used. I have no concerns with this method.

Panel Member #9: Empirical reliability testing assessed both internal inconsistency and test-retest reliability at the individual level, as well as signal-to-noise ratio calculations at the facility level.

7. **Assess the results of reliability testing**

Submission document: Testing attachment, section 2a2.3

Panel Member #1: Results support strong reliability at both the data elements and score levels.

Panel Member #2: Data element: Internal consistency and stability (test-retest) coefficients were consistently high in legacy versions, suggesting same would be true of JR versions.
Measure reliability also high (0.96) with small interquartile range (0.04).

Panel Member #3: The literature indicates the HOOS, Jr. and KOOS, Jr. are sufficiently reliable instruments, including both internal consistency and test-reset reliability.
The signal-to-noise reliability was 0.96 which is in an excellent range.
The exclusion of data element reliability testing for a number of critical data elements is concerning.

Panel Member #4: Performance in the data element testing (for bot instruments: HOOS, JR & KOOS, JR) and measure score reliability testing were strong on all accounts.
"Data Element Reliability
HOOS, JR Reliability:
Internal consistency: … 0.86 in the HSS cohort and 0.87 in the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) cohort….
KOOS, JR Reliability: … were 0.84 in the HSS cohort and 0.85 in the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) cohort.
Measure Score Reliability: … median reliability score of 0.9589 (range: 0.8956 – 0.9916). Interquartile range was 0.0370." [p10]

Panel Member #7: An ICC comparing hospital level results appears not to have been performed. It is therefore difficult to evaluate measure reliability. Further, this appear to be a composite measure (see#2 above).

Panel Member #8: I have no concerns with the results as presented.

Panel Member #9: Clearly acceptable.

Panel Member #10: Assessed both (1) internal reliability and (2) test-test reliability of survey-based outcome measure. These reliability measures were high for THA/TKA. Results from this testing (i.e. test-retest reliability > 0.70) consistent with data element reliability.
Also assessed score-level reliability using SNR which was 0.96 consistent with excellent score-level reliability.

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)

   Panel Member #4: Given there are 19 risk adjustment variables that appear in the risk model, I'd suggest a number of these variables are critical to test, which are primarily the clinical characteristics based on coding (e.g. liver disease, severe infection). While probably not "critical", it would have been desirable to perform testing of select denominator variables.

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

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

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

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

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

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: The 'Low' rating is due to missing information within the specifications, as noted above, in which case NQF guidance is to rate reliability as low.
    Panel Member #2: Although I rated reliability as high, I have two concerns that indirectly can affect reliability. One is the allowance for proxy assessment. This is unorthodox and can add significant noise. It was not clear if proxy results were included in data reported (but I think not). The second concern is missing data…only available complete data analyzed without accounting for what is likely to be fairly extensive missingness. See validity comments.
    Panel Member #3: If we decide that data element reliability testing IS needed in addition to measure score reliability testing, then my recommendation is the testing needs to extend beyond just the HOOS, Jr. and KOOS, Jr. to include all critical data elements.
    Panel Member #4: Response to Q8: Performance in the data element testing (for bot instruments: HOOS, JR & KOOS, JR) and measure score reliability testing were strong on all accounts.
    Response to Q10 [regarding testing of critical data elements]: Given there are 19 risk adjustment variables that appear in the risk model, I'd suggest a number of these variables are critical to test, which are primarily the clinical characteristics based on coding (e.g. liver disease, severe infection). While probably not "critical", it would have been desirable to perform testing of select denominator variables.
    Panel Member #5: Comprehensive assessment of HOOS and KOOS (internal consistency, test retest, and measure score >0.9 for both).
    Panel Member #7: The information provided from the literature is at the patient not hospital level.
    Panel Member #8: I have no concerns with the results as presented.
    Panel Member #9: As the developers note in the application, they used a case-volume cut-off of 25 for the facility-level reliability testing and recommend that this same cutoff be maintained when the measure is in use.

Panel Member #10: Assessed both (1) internal reliability and (2) test-test reliability of survey-based outcome measure. These reliability measures were high for THA/TKA. Results from this testing (i.e. test-retest reliability > 0.70) consistent with data element reliability.
Also assessed score-level reliability using SNR which was 0.96 consistent with excellent score-level reliability.

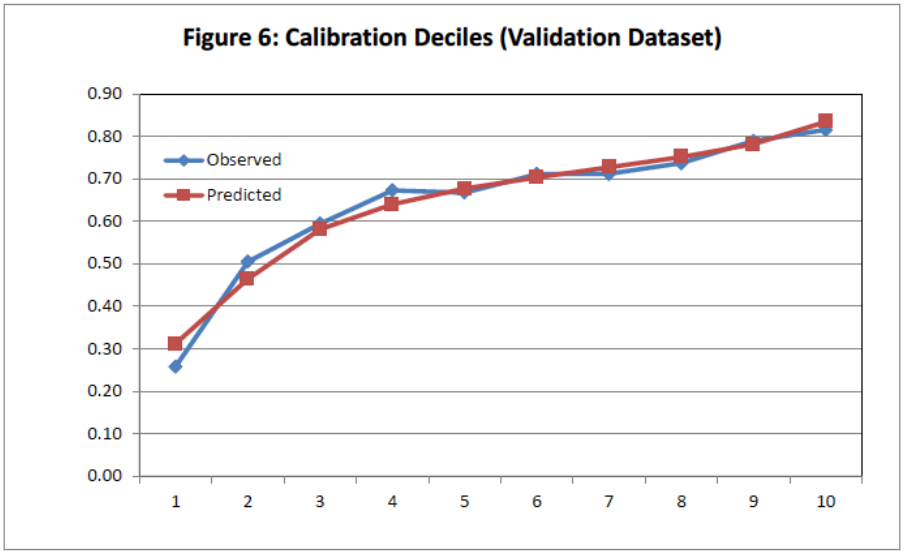
VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

Panel Member #10: Risk adjustment was performed where the outcome was binary 0/1 if patient achieved substantial clinical benefit (SCB) improvement on the PROM score measured as the difference between preoperative and postoperative score. Estimated hospital specific RSIR (risk-standardized improvement ratio based on PE ratio) using hierarchical model. Applied stabilized inverse probability weights (IPW) to address non-response bias.

Cstat in validation data is 0.69, and calibration (-0.08, 1.02) for combined model. Calibration plots also suggest that the model is well calibrated. **These results suggest that model performance is acceptable and support the predictive validity of this measure.**

The risk variables included in the final model are:

- Age, in Years
- Male Sex
- Procedure: THA
- Bilateral procedure
- Health Literacy (assessed by response to Single Item Literacy Screener questionnaire, "Comfort Filling Out Medical Forms by Yourself") (Wallace et al, 2006; Sarkar et al, 20
- Back Pain at preoperative assessment (Quantified Spinal Pain: Patient-Reported Back Oswestry Disability Index question) (Fairbank et al, 2000; Ayers et al, 2013)
- Pain in Non-Operative Lower Extremity Joint (Total painful joint count: Patient-Repor Non-operative Lower Extremity Joint) (Ayers et al, 2013)
- Body Mass Index, in kg per m^2
- Narcotic Use for >90 days
- Baseline PROMIS Global Mental Health Subscale Score
- Severe infection; other infectious diseases (CC 1, 3-7)
- Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)
- Liver disease (CC 27-31)
- Rheumatoid arthritis and inflammatory connective tissue disease (CC 40)
- Depression (CC 61)
- Other psychiatric disorders (CC 63)
- Coronary atherosclerosis or angina (CC 88-89)
- Vascular or circulatory disease (CC 106-109)
- Renal failure (CC 135-140)



We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation is 0.00194 (p-value=0.84). This indicates that there is not an association between the residuals and the probability of response based on our model.

We examined the correlation between the residuals of the stabilized inverse probability weighted hierarchical model and the submission probability finding it to be 0.00492 (p-value=0.60) suggesting that there is not an association between the residuals weighting for non-response and probability of response.

The correlation between RSIR unadjusted and inverse probability weighted RSIR is very high suggesting that the results are not sensitive to our weighting adjustment. However, due to the high proportion of non-responders, we considered it important to account for the differences in characteristics of responders and non-responders found in the literature and empirically in our data. We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation is 0.00194 (p-value=0.84). This indicates that there is not an association between the residuals and the probability of response based on our model.

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

Submission document: Testing attachment, section 2b2.

Panel Member #1: Exclusions do not consider or test for hospitals excluded due to low counts (<25), which represent 52% of hospitals included in the denominator. This is a major potential threat to the measure's validity unless the denominator is redefined as suggested above.

Panel Member #2: Exclusions seem reasonable and well-argued.

Panel Member #3: Measure developer did not provide any data on how the excluded patients impact the performance measure scores.

Panel Member #4: No concerns

Panel Member #7: None

Panel Member #8: Staged procedures are excluded. I understand the theoretical rationale and do not have an issue with this exclusion. However, it might eliminate up to 43% of procedures which seems significant. Perhaps "from unstaged procedures" should be added to the measure name to prevent this from being misleading.

Panel Member #9: No concerns.

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: No concerns. Developers provided evidence for a wide variation in performance between hospitals.
Panel Member #2: Substantial ceiling effects in the data elements (HOOS JR 37-46%, and KOOS JR 19-22%), suggesting inability to produce requires substantial clinical benefit. Many orthopedic surgeons have turned to improved PROMIS measures that greatly reduce floor and ceiling effects. The developers are aware of PROsetta Stone which can be used to link HOOS JR and KOOS JR to PROMIS Pain and Physical Function, addressing this issue and using the PROMIS metric for reporting.
Panel Member #3: No concerns
Panel Member #4: No concerns. Measure results show good variation of hospital performance.
Panel Member #7: It is unclear from the data provided in T.11 whether there are meaningful differences between hospitals in the top quartile. It would be helpful for interpretation to have provided an external validation variable (e.g. 30-day readmission rates for THA/TKA) to aid in interpretation of the magnitude of between hospital differences.

Panel Member #8: I am a bit concerned that there is a ceiling effect for the instruments (esp. for the KOOS, JR). However, this may not be important at this time.

Panel Member #9: No concerns.

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 #3: Not applicable.

Panel Member #4: No concerns

Panel Member #7: N/A

Panel Member #9: No concerns.

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

Submission document: Testing attachment, section 2b6.

Panel Member #1: No concerns. The use of IPW in the risk-adjustment model accounted for potential bias, with the limitation of identifying potential missing data not at random using only the available variables (a common and most probably minor limitation).
Did developers assess the need to adjust for missing data using IPW, or in other words the impact of IPW on hospital ratings by conducting a sensitivity analyses?
Panel Member #2: Extensive missing data threatens the validity of the measure. The proposed solution (stabilized inverse probability weighting), I believe, assumes data are missing at random which is probably not likely. The only good solution here is to require near-complete data rather than rely on proxies and statistical modeling to complete the data table. Excluding deaths seems reasonable given low base rate during the observation period, but there should be a parallel check on death as a possible adverse event (since surgery should have not been done, in hindsight, were death from most other likely causes.
Panel Member #3: It is unclear to me how hospitals with < 25 cases with complete PRO data are treated. The measure developers did not provide any data on how hospitals with < 25 cases compare to those that >=25 cases.

Panel Member #4: The non-response rate is concerning at 43%. However, the testing for bias in non-respondents (vs respondents) based on clinical characteristics appears preliminarily that the low response rate does not necessarily introduce a bias in the results / measure ratings.
"The true "response" rate for our study is difficult to calculate because it is unknown to whether 100% of eligible patients at the hospitals in our dataset were asked to provide PRO data… [M]ean response rate among hospitals was 43.15% (See Table 13, below)." [p37]
"We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation is 0.00194 (p-value=0.84). This indicates that there is not an association between the residuals and the probability of response based on our model." [p39]
Panel Member #7: It appears from the data provided that more than 50% of hospitals with >25 eligible patients and about 70% of all hospitals had missing PRO and/or risk variable data. Despite the proposed propensity score analysis used to access comparable of those with and without missing data, there remains concerns about substantial bias in the data presented. Further, it appears that "missing" was applied to the pre-op, post-op PRO surveys vs. missing responses within the survey which could introduce additional bias and was not mentioned in section 2b6.

Panel Member #8: Stabilized inverse probability weighting was used to assess the impact of non-response bias. The measure developers concluded that the correlation between the unweighted and weighted measures was very high, thus indicating that results were not sensitive to adjustment. I agree with this conclusion and have no concerns with the methods used.

Panel Member #9: No concerns.

16. Risk Adjustment

Panel Member #2: 19 risk factors in model

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:**

Panel Member #10: Health literacy, which was associated with better outcomes was included in outcome. This does adjust for the fact that lower SES patients, who may have lower health literacy, will have worse outcomes.

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

Panel Member #1: health literacy
Panel Member #4: Health literacy included in the risk model

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

Panel Member #4: See response to 2b3.4b [p27]

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

Panel Member #2: Retained only health literacy…probably surrogate for other social factors. But how will it be measured in practice?
Panel Member #4: Health literacy included in the risk model. See response to 2b3.4b [p27]

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

Panel Member #4: NA – present at onset

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 overall development process and testing of the risk-adjustment approach seems solid.

I do have several concerns/questions for clarity:

- The model was developed including cases from hospitals not used for reliability, validity and missing data testing, i.e., hospitals with low caseloads (n<25) not recommended for this measure. Did the developers do a sensitivity test to assess the impact of excluding these hospitals from the risk-adjustment development sample on the risk-adjustment model?
- The hospital performance outcome (Risk-Standardized-Improvement-Rate: RSIR) is calculated as the ratio of predicted and expected improvement, multiplied by the overall observed improvement rate. The latter (overall observed improvement rate) is not defined, or I might have missed it somewhere in the submission. I am assuming the overall observed improvement rate is calculated as the overall average of 0 & 1 for substantial clinical benefit (SCB) for all patients from all hospitals. Can the developers clarify if this is in fact what is meant by overall observed improvement rate, and if so, why would this rate be used both in the development of the HLM as the dependent variable, and then again in the calculation of the RSIR? Couldn't this contribute to some circularity in the calculation of RSIR? A practical example of the calculation of a hospital specific RSIR would be very helpful to clarify and demonstrate how the actual calculation is conducted. This information would be good to have in the measure specification as well as in the risk-adjustment section of the testing form.
- I would like to better understand the logic in selecting a single threshold for SCB (by THA/TKA). There is a wealth of published literature on the dependency of clinically important improvement thresholds on initial scores. It is common to see that patients with worse initial scores need more positive change points to reach clinically important improvement thresholds, compared to patients with high initial scores. Without considering admission status, in this case pre-surgical status, SCB values could potentially over-estimate or under-estimate the intended outcome. Developers noted that "the SCB outcome allows patients with poor baseline PRO scores to improve, so some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB.". Doesn't this bias the SCB measure into having better outcomes compared to those estimated by a patient with low admission scores who might expect more change for it to be substantial? The developer's note above seems problematic since an opposite case would also occur, that is a patient with higher admission scores that might have lower expectations for substantial improvement, yet they would have a lesser chance of achieving the predetermined SCB threshold, which would penalize providers with higher performing patients at admission.
- Finally, could developers explain why they chose to include in the risk-adjustment model multiple factors that were not significant?

Panel Member #3: Concerns with the lack of adjustment for non-English speakers, given that the KOOS, Jr. and HOOS, Jr. are only offered in English

Good discrimination: c-statistic of 0.68, 0.69

Panel Member #4: Reasonable and appropriate method to assess the risk model. However, the testing results in regard to the c-statistic are marginally acceptable: 0.68 (development data) and 0.69 (validation data).

Testing method:

"To assess Model Performance, we computed discrimination and calibration statistics for assessing model performance (Harrell and Shih, 2001) for the clinically derived models, including:

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

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; good discrimination indicated by a wide range between the lowest decile and highest decile); and

(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). A value of close to zero for the intercept and close to 1 for coefficient of risk score indicates good calibration of the model." [p29]

Testing results:

For the Development Dataset:

- C-statistic for the risk model is 0.68
- Predictive ability from the lowest to highest decile is 26% - 82%

For the Validation Dataset:

- C-statistic for the risk model is 0.69
- Predictive ability from the lowest to highest decile is 26% - 81%" [p30]

Panel Member #5: Standard multivariable risk adjustment, includes health literacy. C-stat 0.68.

Panel Member #7: From the data provided, it appears that the risk adjustments strategy is adequate.

Panel Member #8: I am a bit concerned that the only social risk factor included is health literacy level

Panel Member #9: Thoughtful and appropriate.

Panel Member #10: Risk adjustment was performed where the outcome was binary 0/1 if patient achieved substantial clinical benefit (SCB) improvement on the PROM score measured as the difference between preoperative and postoperative score. Estimated hospital specific RSIR (risk-standardized improvement ratio based on PE ratio) using hierarchical model. Applied stabilized inverse probability weights (IPW) to address non-response bias.

Cstat in validation data is 0.69, and calibration (-0.08, 1.02) for combined model. Calibration plots also suggest that the model is well calibrated. These results suggest that model performance is acceptable and support the predictive validity of this measure.

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

Panel Member #10:

HOOS, JR Validity:

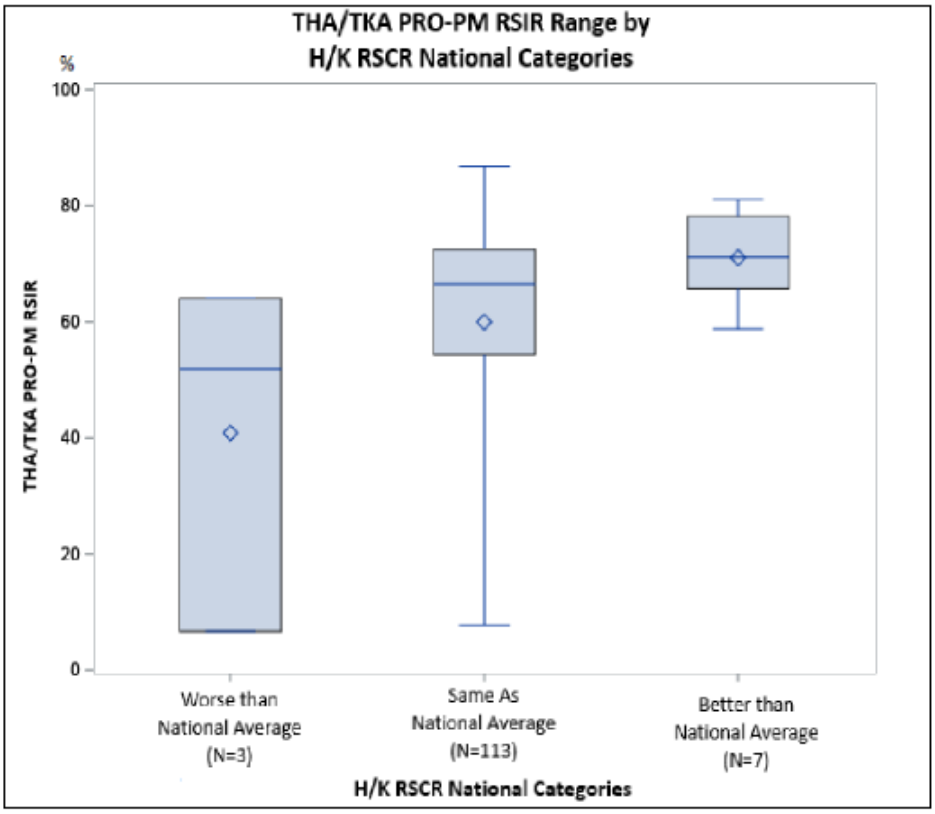
Responsiveness: Responsiveness of the HOOS, JR to changes following a total hip replacement was evaluated using standardized response means, and then examined against other previously validated PROMs (HOOS domains, The Western Ontario and McMaster University Arthritis Index [WOMAC] domains) in the HSS cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) registry at 2 years after a THA procedure (Lyman et al, 2016a). A standardized response mean greater than 0.8 was considered large (Steiner and Norman, 2003).

External validity: External construct validity was evaluated using Spearman's correlations between HOOS, JR and the HOOS and the WOMAC. A Spearman's correlation coefficient of 0.8 or greater was considered very high external validity (Wechsler, 1996). External correlations were assessed using a scatterplot overlying a contour plot based on bivariate kernel density estimation between the HOOS, JR and HOOS domains (Lyman et al, 2016a).

Floor and ceiling effects: Floor and ceiling effects (percent at worst possible score preoperatively and best possible score postoperatively) were evaluated against the HOOS and the WOMAC instruments (Lyman et al, 2016a).

The validity of the outcome data elements were tested using measures of responsiveness and external validity. These testing appear to indicate that the survey outcome data elements are valid.

The measures were also validated by examining the PROM in low, average and high performing hospitals identified using risk-standardized complication raters. These analyses suggest that higher performing hospitals, based on their complication rates, also have better rates of higher functional outcomes:



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

Submission document: Testing attachment, section

Panel Member #1: I have no concerns with the validity testing methods used.
Panel Member #2: Expert panel (face) and data element validity relied on PRO developer data and publications. Measure validity accomplished by comparing risk-standardized improvement rates (RSIRs) to NQF#1550 data: Hip/Knee Complication Measure. Creates 3 groups: worse; same; better. focus on hospitals submitting COMPLETE data on >25 procedures.
Panel Member #3: Provided information on validity testing for the HOOS, Jr. and KOOS, Jr., but did not provide any data/analysis of the validity testing of other critical data elements.
For empirical validity testing, compared hospital performance to complication rate measure, with the hypothesis that these measures would move in the same direction
Panel Member #4: The construct validity testing appears to be appropriate for measure score testing. "To assess empirical measure score validity we compared the THA/TKA PRO-PM risk-standardized improvement rates (RSIRs) to the NQF endorsed Hip/Knee Complication Measure (NQF #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA.)" [p13]
Panel Member #7: The data element validity presented refers to studies in the literature, is at the patient level and is not presents in the units constituting cut-offs described in the measure specifications.
Panel Member #8: Data element validity assessments included: responsiveness using standardized response means (which I am not sure I understand - does this mean the mean of the patient scores?), evaluated against other PROMS; external validity using spearman's correlations between the HOOS/HOOS and the HOOS/KOOS, JR; and floor and ceiling effects.
Empirical measure score validity assessments included comparing the THA/TKA PRO-PM risk-standardized improvement rates (RSIRs) to the NQF endorsed Hip/Knee Complication Measure (NQF #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA.)
I find all of these methods to be scientifically acceptable.
Panel Member #9: Data-element level validity testing was conducted appropriately for both scales by comparing scales and individual questions to one another. In addition, validity testing was conducted at the measure level by comparing the HOOS, JR and the KOOS, JR PROM instruments across hospitals.
Panel Member #10: See above

22. Assess the results(s) for establishing validity

Submission document: Testing attachment, section 2b2

Panel Member #1: Validity of data elements are overall supportive of high levels of validity, except the substantial ceiling effect of both the HOOS (37-46%) and the KOOS (19-22%). However, similar ceiling effects of these measures have been reported preciously and are not specific to this patient population. Also, the binary use of the HOOS JR and KOOS JR scores might mitigate concerns regarding ceiling effects. Having said that, recent publications have supported the use of other non-condition specific measures (e.g. PROMIS physical function) as valid alternatives to the HOOS & KOOS that may reduce these large ceiling effects and improve responsiveness (e.g., Hung et al 2018; Padilla et al 2019; Kortlever 2019; Kenney 2019; Li et al 2020). Some consideration to these alternatives may be of value for future consideration by the developers.

Panel Member #2: (P16) large differences between NQF #1550 groups on data element. Overall, relative to what I would expect, few patients appear to report substantial clinical improvement. This could be because the bar is set too high, or ceiling effects of the measures, or both.
Panel Member #3: Literature shows that HOOS, Jr. and KOOS, Jr. are valid instruments; no data on other critical data elements
Found a general trend of better performance on the complication measure reflected better performance on the PRO measure.
Panel Member #4: The construct validity testing suggests there is a modest (and expected) relationship between this measure and the NQF endorsed hip & knee complication measure.
"…hospitals in the "RSCR Worse than National Average" category have lower median RSIRs (51.87%) than the median RSIR (66.49%) of hospitals in the "RSCR Same as National Average" category, which is lower than that of hospitals in the "RSCR Better than National Average" category (71.13%). The hospitals with lower risk-adjusted complication rates have higher risk-adjusted THA/TKA improvement rates (Figure 1)." [p15]
Panel Member #7: Although the association of the THA/TKA PRO-PM RSIR with the hospital risk standardized complication rate (NQF: 1550) was described in the text as a "correlation" those data were not presents. Data in Figure 1 display box plots (the analysis generating these data were not clear) with evidence of considerable validity in results at the mean. A plot of the association of pass/fail on each measure at the hospital level would have been helpful.
Panel Member #8: Data element validity: responsiveness (no concerns); external validity (no concerns); and floor and ceiling effects (it seems like the HOOS has a ceiling effect of between 37% and 46%. This seems really large to me).
Empirical measure score validity: (no concerns)
Panel Member #9: The results are very strong.

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 <u>are</u> threats to validity and/or relevant threats to validity were <u>not assessed OR</u> 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: Validity testing was appropriate with results supporting the overall validity of this measure at both the data elements and score level.
However, there are threats to validity to be considered as noted above, which is why I rated validity as insufficient with the intention of receiving additional information that might alleviate these concerns:

- Exclusion of roughly half of the included hospitals in validity testing. A revision of the measure specifications could easily address this issue.
- Concerns and need for clarifications that are detailed above regarding the risk-adjustment model.

Panel Member #2: On a purely procedural level, based upon the activities done leading to the presentation of this measure, one could assess it as having moderate validity, but the following factors led me to come down on the side of Low validity. Missing data concerns, proxy reporting allowance, ceiling effects of data element, and missed opportunity to enable practices more choice of data element capture (i.e., PROMIS and PROsetta Stone).
Panel Member #3: If we decide that data element validity testing IS needed in addition to measure score validity testing, then my recommendation is the testing needs to extend beyond just the HOOS, Jr. and KOOS, Jr. to include all critical data elements.
As noted above, there are also concerns with better understanding excluded patients and missing data
Panel Member #4: Response to 16e: … the testing results in regard to the c-statistic are marginally acceptable: 0.68 (development data) and 0.69 (validation data).
Response to Q23: The construct validity testing suggests there is a modest (and expected) relationship between this measure and the NQF endorsed hip & knee complication measure.
Panel Member #5: Empiric validity testing and systematic assessment of face validity. Justification of binary outcome provided.
Panel Member #7: Substantial amount of missing data, even for hospitals with ≥ 25 eligible patients, raises concerns about bias not adequately addressed in the proposal.
Panel Member #9: Data-element level validity testing was conducted appropriately for both scales by comparing scales and individual questions to one another. In addition, validity testing was conducted at the measure level by comparing the HOOS, JR and the KOOS, JR PROM instruments across hospitals. Both sets of results were strong.
Panel Member #10: Performance of risk adjustment model consistent with acceptable level of predictive validity:
Risk adjustment was performed where the outcome was binary 0/1 if patient achieved substantial clinical benefit (SCB) improvement on the PROM score measured as the difference between preoperative and postoperative score. Estimated hospital specific RSIR (risk-standardized improvement ratio based on PE ratio) using hierarchical model. Applied stabilized inverse probability weights (IPW) to address non-response bias.
Cstat in validation data is 0.69, and calibration (-0.08, 1.02) for combined model. Calibration plots also suggest that the model is well calibrated. These results suggest that model performance is acceptable and support the predictive validity of this measure.
Evaluation of outcome data elements consistent with acceptable level of validity at the data element level: The validity of the outcome data elements were tested using measures of responsiveness and external validity. These testing appear to indicate that the survey outcome data elements are valid.

The measures were also validated by examining the PROM in low, average and high performing hospitals identified using risk-standardized complication raters. These analyses suggest that higher performing hospitals, based on their complication rates, also have better rates of higher functional outcomes:

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

Panel Member #7: No data on composite measure or the use of scores for such measures at the hospital level provided.

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.

Panel Member #1: Clarifications to the specifications and risk adjustment method might address the concerns raised above, in which case further discussion might not be needed.

Panel Member #8: I found this application difficult to read (and frustrating) because the creation of the instrument was never discussed. I wanted to see that the FDA guidance for PROs were followed and I had to dig in articles and on the internet for quite a while before I found that the KOOS/HOOS, JRs (which were based on the KOOS/HOOS was derived from the WOMAC = Western Ontario and McMaster Universities Osteoarthritis Index which was created in 1982. That was as far back as I could trace the instrument and thus, I know nothing of how the instrument was created (i.e., were focus groups completed until concept saturation was reached during the creation of the initial instrument, etc.).

Also, I would have liked to have seen evidence of the content coverage (content validity) for each measure.

Developer Submission

Additional evaluations and submission materials attachments…

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

Del4-7b2HBPNQFHipKneePROPMEvidNQFForm_3.5.20.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.

No

red font to indicate updated evidence.

1a. Evidence (subcriterion 1a)

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

Measure Title: Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

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

Date of Submission: TBD

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

Outcome

☒ Outcome: Click here to name the health outcome

☒ Patient-reported outcome (PRO): Click here to name the PRO

PROs include HRQoL/functional status, symptom/symptom burden, experience with care, health-related behaviors. (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*): Click here to name the intermediate outcome

☐ Process: Click here to name what is being measured

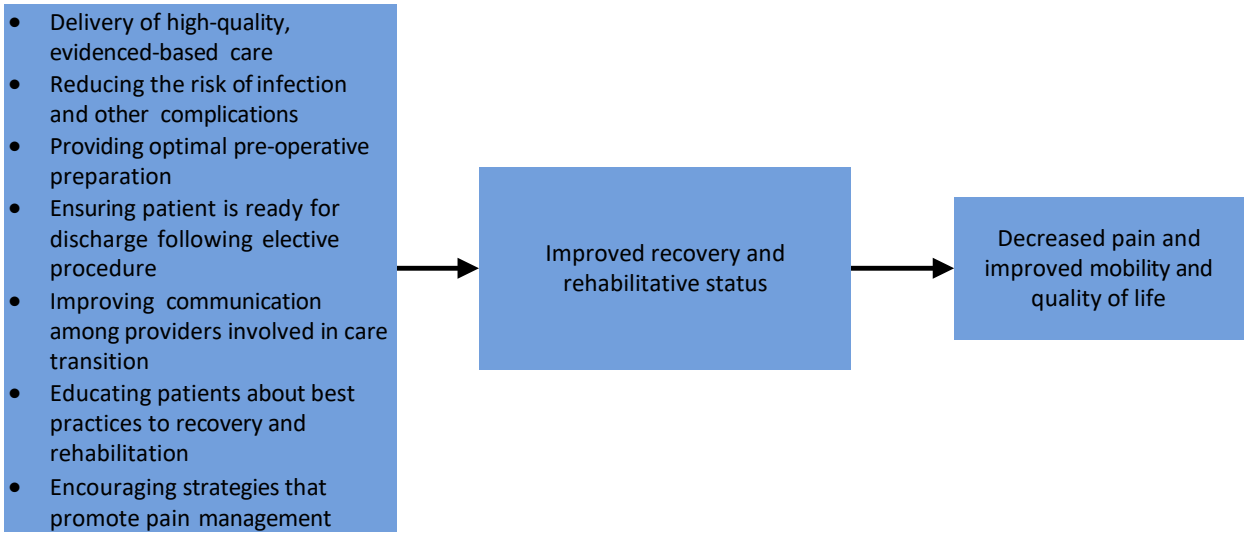
☐ Appropriate use measure: Click here to name what is being measured

☐ Structure: Click here to name the structure

☐ Composite: Click here to name what is being measured

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.



The goal of this measure is to directly affect patient outcomes by measuring patient-reported outcomes following total hip and/or total knee arthroplasty (THA/TKA). Measurement of patient-reported outcomes, including pain and functional status, allows for a broad view of quality of care that encompasses more than what can be captured by individual process-of-care measures. More specifically, functional status following THA/TKA is likely to be influenced by a broad range of clinical activities such as prevention of complications and provision of evidenced-based care. The patient is the most appropriate source for such information, and patients have identified that the information that will be captured by this outcome measure is important (Liebs et al., 2013).

References:

Liebs TR, Herzberg W, Gluth J, et al. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. *Bone Joint J.* 2013; 95-B: 239–43.

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.)

Patients who have undergone a THA or TKA have been engaged for input on measure development through participation on the Technical Expert Panel (TEP) and through a Patient Working Group assembled with assistance from the National Partnership for Women and Families in 2018. Overall, five patients (two males and three females) have provided input through TEP participation: two patients participated in four TEP meetings in 2013 and 2014; they were unavailable to continue participation when the TEP was reconvened in 2018, and two new patients participated in two TEP meetings in 2018 and 2019; and a fifth patient participated in the final TEP meeting in 2020 when one of two prior patients could not continue. The Patient Working Group consisted of five females and one male who have undergone at least one hip and/or knee replacement and were distinct from those who participated in the TEP. These patients were convened for three meetings, one in July 2018, one in February 2019, and one in February 2020. Additional input was sought from both the TEP and the Patient Working Group through online surveys following some of their meetings.

Feedback from patients on both the TEP and the Patient Working Group indicate strong support for a patient-reported outcomes-based performance measure following primary elective THA and TKA. Patients stated that they expect a significant amount of improvement in both pain level and functional status following a THA/TKA procedure and felt this was an extremely important aspect of care to be captured in this measure. Patients also noted that their surgical experience positively impacted not only their physical health, but their quality of life as well. Patients in the Patient Working Group supported a measure cohort that combined THA and TKA patients, while two patients on the TEP expressed some concern about differing postoperative recovery for hips and knees. All patients supported the risk model and accounting for social risk factors in an analytic approach to non-response bias. Patients expressed a desire to see measure results that reflect physician-level performance but agreed that a hospital-level measure is a good way to encourage communication across providers to improve coordination of care at a facility overall.

****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.

THA/TKAs are important, effective procedures performed on a broad population, and the patient-reported outcomes for these procedures (for example, pain, mobility, and quality of life) can be measured in a scientifically sound way (Alviar et al., 2011 [a]; Alviar et al., 2011 [b]; Bauman et al., 2007; Collins & Roos, 2012; Jones et al., 2007; Jones & Pohar, 2012; Lau et al., 2012; Liebs, 2016; Montin et al., 2008; Papalia et al., 2012; Rolfson et al., 2011; Thorborg et al., 2010; White & Master, 2016) and are influenced by a range of improvements across the full spectrum of care. THA/TKA provides a suitable environment for optimizing care, as there are many studies indicating how providers can improve outcomes of the patients by addressing aspects of pre-, peri-, and postoperative care (Brown et al., 2012; Choong et al., 2009; Galea et al., 2008; Kim, 2019; McGregor et al., 2004; Moffet et al., 2004; Monticone et al., 2013; Walters, 2016).

Optimal clinical outcomes depend not just on the surgeon performing the procedure, but also on: the entirety of the team's efforts in the care of the patient; care coordination across provider groups and specialties; and the patients' engagement in their recovery (Feng et al, 2018; Saufl et al, 2007). Even the best surgeon will not get outstanding results if there are gaps in the quality of care provided by others caring for the patient before, during, and/or after surgery. The goal of hospital-level outcome measurement is to capture the full spectrum of care to incentivize collaboration and shared responsibility for improving patients' health and reducing the burden of their disease.

References:

Alviar M, Olver J, Brand C, Hale T, Khan F. Do Patient-Reported Outcome Measures Used in Assessing Outcomes in Rehabilitation After Hip and Knee Arthroplasty Capture Issues Relevant to Patients? Results of a Systematic Review and ICF Linking Process. *J Rehabil Med.* 2011; 43:374-381. [a]

Alviar M, Olver J, Brand C, et al. Do Patient-Reported Outcome Measures in Hip and Knee Arthroplasty Rehabilitation Have Robust Measurement Attributes? A Systematic Review. *J Rehabil Med.* 2011; 43:572-583. [b]

Bauman S, Williams D, Petruccelli D, Elliott W, de Beer J. Physical Activity After Total Joint Replacement: A Cross-Sectional Survey. *Clin J Sport Med.* 2007; 17(2):104-108.

Brown K, Topp R, Brosky JA, Lajoie AS. Prehabilitation and quality of life three months after total knee arthroplasty: a pilot study. *Percept Mot Skills.* Dec 2012; 115(3):765-774.

Choong PF, Dowsey MM, Stoney JD. Does accurate anatomical alignment result in better function and quality of life? Comparing conventional and computer-assisted total knee arthroplasty. *J Arthroplasty.* Jun 2009; 24(4):560-569.

Collins NJ, Roos EM. Patient-reported outcomes for total hip and knee arthroplasty: commonly used instruments and attributes of a "good" measure. *Clin Geriatr Med.* 2012; 28(3):367-394.

Feng JE, Novikov D, Anoushiravanni AA, Schwarzkopf R. Total knee arthroplasty: Improving outcomes with a multidisciplinary approach. *J Multidiscip Healthc.* 2018; 11:63-73. doi: 10.2147/JMDH.S140550.

Galea MP, Levinger P, Lythgo N, et al. A targeted home-and center-based exercise program for people after total hip replacement: a randomized clinical trial. *Arch Phys Med Rehabil.* Aug 2008; 89(8):1442-1447.

Jones CA, Beaupre LA, Johnston DW, Suarez-Almazor ME. Total joint arthroplasties: current concepts of patient outcomes after surgery. *Rheum Dis Clin North Am.* 2007; 33(1):71-86.

Jones CA, Pohar S. Health-related quality of life after total joint arthroplasty: a scoping review. *Clin Geriatr Med.* 2012; 28(3):395-429.

Kim KY. Perioperative orthopedic surgical home: Optimizing total joint arthroplasty candidates and preventing readmission. *J Arthroplasty.* 2019; 34(7s):S91-S96. doi: 10.1016/j/arth.2019.01.020.

Lau RL, Gandhi R, Mahomed S, Mahomed N. Patient satisfaction after total knee and hip arthroplasty. *Clin Geriatr Med.* 2012; 28(3):349-365.

Liebs TR. Quality-adjusted life years gained by hip and knee replacement surgery and its aftercare. *Arch Physical Med Rehabil.* 2016; 97(5):691-700. doi: 10.1016/j.apmr.2015.12.021.

McGregor AH, Rylands H, Owen A, Dore CJ, Hughes SP. Does preoperative hip rehabilitation advice improve recovery and patient satisfaction? *J Arthroplasty.* Jun 2004; 19(4):464-468.

Moffet H, Collet JP, Shapiro SH, Paradis G, Marquis F, Roy L. Effectiveness of intensive rehabilitation on functional ability and quality of life after first total knee arthroplasty: A single-blind randomized controlled trial. *Arch Phys Med Rehabil.* Apr 2004; 85(4):546-556.

Monticone M, Ferrante S, Rocca B, et al. Home-based functional exercises aimed at managing kinesiophobia contribute to improving disability and quality of life of patients undergoing total knee arthroplasty: a randomized controlled trial. *Arch Phys Med Rehabil.* Feb 2013; 94(2):231-239.

Montin L, Leino-Kilpi H, Suominen T, Lepisto J. A systematic review of empirical studies between 1966 and 2005 of patient outcomes of total hip arthroplasty and related factors. *J Clin Nurs.* 2008; 17(1):40-45.

Papalia R, Del Buono A, Zampogna B, Maffulli N, Denaro V. Sport activity following joint arthroplasty: a systematic review. *Br Med Bull.* 2012; 101:81-103.

Rolfson O, Rothwell A, Sedrakyan A, et al. Use of patient-reported outcomes in the context of different levels of data. *J Bone Joint Surg Am.* 2011; 3:66-71.

Saufl N, Owens A, Kelly I, Merrill B, Freyaldenhouen L. A multidisciplinary approach to total joint replacement. *J Perianesth Nurs.* 2007; 22(3):195-206.e9.

Thorborg K, Roos EM, Bartels EM, Petersen J, Holmich P. Validity, reliability and responsiveness of patient-reported outcome questionnaires when assessing hip and groin disability: a systematic review. *British Journal of Sports Medicine.* 2010; 44(16):1186-1196.

Walters M. Reducing length of stay in total joint arthroplasty care. *Orthop Clin North Am.* 2016; 47(4):653-660. doi: 10.1016/j.ocl.2016.05.006.

White DK, Master H. Patient-reported measures of physical function in knee osteoarthritis. *Rheum Dis Clin North Am.*2016; 42(2):239-352. doi: 10.1016/j.rdc.2016.01.005.

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

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

Source of Systematic Review: • **Title** • **Author** • **Date** • **Citation, including page number** • **URL**	**N/A.**
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?	

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 information to patients, physicians, and hospitals about hospital-level, risk-standardized patient-reported outcomes, such as pain and functional status, following elective primary THA/TKA. Measurement of patient-reported outcomes allows for a broad view of quality of care. Complex and critical aspects of care — such as communication among 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-of-care measures. As patient outcomes are not only influenced by care given during the time of hospitalization but also by patient status on presentation, outcome measures ideally are risk adjusted for patients' comorbid conditions.

THA/TKA procedures provide a particularly rich test bed for developing quality measures based upon patient-reported experiences. These procedures are commonly performed in older patients who have marked pain and functional limitation preoperatively, and who often experience significant improvements postoperatively. Patients who have undergone THA/TKA procedures have already indicated their support of such outcomes in the published literature (Liebs et al., 2013) and voiced their support for this measure as part of a TEP and a Patient Working Group.

References:

Liebs TR, Herzberg W, Gluth J, et al. Using the patient's perspective to develop function short forms specific to total hip and knee replacement based on WOMAC function items. Bone Joint J. 2013; 95-B:239–43

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.*

Table 1. Mean and Distribution of Hospital-level Risk Standardized Improvement Rates (RSIRs) for Risk Model for SCB Improvement following Elective Primary THA/TKA Performed July 1, 2016 to June 30, 2017 (Hospitals with >25 THA/TKA Patients with PRO Data)

Statistic// THA/TKA Procedures

N (Hospitals)// 123

Mean (SD)// 60.16% (219.58)

Percentile

100% Max// 86.84%

99%// 84.73%

95%// 81.92%

90%// 78.85%

75% (Q3)// 72.51%

50% (Median)// 66.49%

25% (Q1)// 54.36%

10%// 20.94%

5%// 13.42%

1%// 7.70%

0%// 6.65%

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.

As stated previously, THA/TKA procedures are commonly performed in older patients who have marked pain and functional limitation preoperatively, and who often experience significant improvements postoperatively. However, not all patients experience benefit from THA/TKA procedures (National Joint Registry, 2012), and many note that their preoperative expectations for functional improvement have not been met (Ghomrawi et al., 2011; Harris et al., 2013; Jourdan et al., 2012; Suda et al., 2010). While the degree and extent of variation in these outcomes across hospitals in the U.S. is unknown, variation in clinical practice has been well documented in the U.S. (American Academy of Orthopedic Surgeons, 2011; Anderson et al., 2012; Roos, 2003). Readmission and complication rates vary across hospitals (Suter et al., 2013a; Suter et al., 2013b), and international experience documents hospital-level variation in patient-reported outcome measures following THA/TKA. The United Kingdom data demonstrated greater than 15% differences among hospitals in the proportion of patients who improved after surgery (National Health System, 2012; Neuburger et al., 2013); and THA/TKA surgical practices vary broadly (American Academy of Orthopaedic Surgeons, 2011; Anderson et al., 2012). Data from this measure support high variability in hospital performance, as noted above.

References:

American Academy of Orthopaedic Surgeons (AAOS). Preventing Venous Thromboembolic Disease in Patients Undergoing Elective Hip and Knee Arthroplasty: Evidence-Based Guideline and Evidence Report. 2011.

Anderson FA, Jr., Huang W, Friedman RJ, Kwong LM, Lieberman JR, Pellegrini VD, Jr. Prevention of venous thromboembolism after hip or knee arthroplasty: findings from a 2008 survey of US orthopedic surgeons. The Journal of arthroplasty. May 2012; 27(5):659-666 e655.

Ghomrawi HM, Franco Ferrando N, Mandl LA, Do H, Noor N, Gonzalez Della Valle A. How Often are Patient and Surgeon Recovery Expectations for Total Joint Arthroplasty Aligned? Results of a Pilot Study. HSS journal: the musculoskeletal journal of Hospital for Special Surgery. Oct 2011; 7(3):229-234.

Harris IA, Harris AM, Naylor JM, Adie S, Mittal R, Dao AT. Discordance between patient and surgeon satisfaction after total joint arthroplasty. The Journal of arthroplasty. May 2013; 28(5):722-727.

Jourdan C, Poiraudeau S, Descamps S, et al. Comparison of patient and surgeon expectations of total hip arthroplasty. PloS one. 2012; 7(1):e30195.

National Health System: The Information Centre for Health and Social Care. HESonline Hospital Episode Statistics: Proms Data. http://www.hesonline.nhs.uk/Ease/ContentServer?siteID=1937&categoryID=1295, 2012.

National Joint Registry. National Joint Registry for England and Wales 9th Annual Report 2012. Available at www.njrcentre.org.uk: National Joint Registry; 2012.

Neuburger J, Hutchings A, van der Meulen J, Black N. Using patient-reported outcomes (PROs) to compare the providers of surgery: does the choice of measure matter? Medical Care. Jun 2013; 51(6):517-523.

Roos EM. Effectiveness and practice variation of rehabilitation after joint replacement. Current opinion in rheumatology. Mar 2003; 15(2):160-162.

Suda AJ, Seeger JB, Bitsch RG, Krueger M, Clarius M. Are patients´ expectations of hip and knee arthroplasty fulfilled? A prospective study of 130 patients. Orthopedics. Feb 2010; 33(2):76-80.

Suter LG, Grady JN, Lin Z, et al. 2013 Measure Updates and Specifications: Elective Primary Total Hip Arthroplasty (THA) And/OR Total Knee Arthroplasty (TKA) All-Cause Unplanned 30-Day Risk-Standardized Readmission Measure (Version 2.0). March 2013a; Available at: http://qualitynet.org/.

Suter LG, Parzynski CS, Grady JN, et al. 2013 Measures Update and Specifications: Elective Primary Total Hip Arthroplasty (THA) AND/OR Total Knee Arthroplasty (TKA) Risk-Standardized Complication Measure (Version 2.0). March 2013b; Available at: http://qualitynet.org/.

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.*

Disparities data are presented below for the Development Dataset (n=6734 patients). These data show bivariate and multivariate results for the following social risk factors: race (non-White), insurance status (Dual eligibility: Medicare and Medicaid coverage), and socioeconomic status (AHRQ SES Index: Bottom quartile). Chi-square analyses and multivariate analyses reveal no statistically significant association between non-White race or AHRQ SES Index bottom quartile and SCB improvement in our Development Dataset; dual eligibility was borderline significant (p=0.058) at the bivariate level (see Table 2 below), and statistically significant when entered into the risk model (see Table 3 below).

Table 2. Bivariate Associations between Social Risk Factors and Observed SCB Improvement (Development Dataset, N=6734)

Variable // Total: Frequency (%) // Achieved SCB Improvement: Frequency (%) // Did Not Achieve SCB Improvement: Frequency (%) // P-value

Race: Non-White // 548 (8.14%) // 351 (8.06%) // 197 (8.28%) // 0.7569

Dual eligibility: Medicare and Medicaid // 206 (3.06%) // 146 (3.35%) // 60 (2.52%) // 0.0580

AHRQ SES Index: Bottom quartile // 688 (10.22%) // 446 (10.24%) // 242 (10.17%) // 0.9922

Table 3. Adjusted Odds Ratios (ORs) for Social Risk Factors Individually Evaluated in the Risk Model for SCB Improvement (Development Dataset, N=6734)

Variable // Frequency (%) // Estimate (SE) // OR (95% CI) // C Statistic for Model Including Social Risk Factor

Race: Non-White // 548 (8.14%) // -0.08 (0.10) // 0.93 (0.76, 1.13) // 0.68*

Dual eligibility: Medicare and Medicaid // 206 (3.06%) // 0.40 (0.17) // 1.49 (1.07, 2.08) // 0.68*

AHRQ SES Index: Bottom quartile // 688 (10.22%) // 0.04 (0.09) // 1.04 (0.87, 1.25) // 0.68*

* C-statistic for the risk model for SCB improvement in the Development Dataset without any of the three social risk factors = 0.68

Table 4. Mean and Distribution of Risk Standardized Improvement Rates (RSIRs) Calculated without and with Social Risk Factors in the Risk Model (Development Dataset: Hospitals with >25 THA/TKA Patients with PRO Data)

Statistic // No Social Risk Factors Included // Race: Non-White // Dual Eligibility // AHRQ SES Index: Bottom Quartile

N (Hospitals) // 94 // 94 // 94 // 94

Mean (SD) // 60.39% (19.85) // 60.36% (19.87) // 60.40% (19.85) // 60.30% (19.86)

Percentile

100% Max // 86.25% // 86.03% // 86.21% // 86.23%

99% // 86.25% // 86.03% // 86.21% // 86.23%

95% // 81.94% // 81.71% // 81.96% // 82.03%

90% // 79.95% // 80.10% // 79.95% // 79.95%

75% (Q3) // 72.37% // 72.45% // 72.38% // 72.33%

50% (Median) // 66.57% // 66.60% // 66.53% // 66.57%

25% (Q1) // 53.22% // 53.26% // 53.23% // 53.22%

10% // 20.07% // 20.04% // 20.08% // 20.06%

5% // 14.47% // 14.43% // 14.49% // 14.50%

1% // 8.47% // 8.42% // 8.48% // 8.46%

0% // 8.47% // 8.42% // 8.48% // 8.46%

Pearson Correlation Coefficient (in association with "No Social Risk Factors")

Race: Non-White: 0.9997

Dual Eligibility: 0.9999

AHRQ SES Index: Bottom Quartile: >0.9999

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):*

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

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

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

N/A

S.2a. <u>If this is an eMeasure</u>, 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**:** HipKneePROPMTestNQFForm_For_Submission_Updated_1-30-2020-637160783322301683.docx

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.

Attachment **:** HipKneePROPMInstruments_For_Submission-637160780855757257.xlsx

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.

Patient

S.3.1. <u>For maintenance of endorsement:</u> 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. <u>For maintenance of endorsement,</u> please briefly describe any important changes to the measure specifications since last measure update and explain the reasons.

N/A

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

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

The numerator is the risk-standardized proportion of patients undergoing an elective primary THA or TKA who meet or exceed an a priori, patient-defined substantial clinical benefit (SCB) threshold of improvement between preoperative and postoperative assessments on joint-specific patient-reported outcome measure (PROM) surveys. SCB improvement is defined as follows:

- For THA patients, an increase of 22 points or more on the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR); and

- For TKA patients, an increase of 20 points or more on the Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR).

SCB thresholds were defined using published literature (Lyman and Lee, 2018) and vetted by our Patient Working Group, Technical Expert Panel (TEP) and Technical Advisory Group.

References:

Lyman S and Lee YY. (2018). What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? Clin Orthop Relat Res, 467(12):2432-2441.

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).

This is a patient-reported outcome-based performance measure (PRO-PM).

Two joint-specific Patient Reported Outcome Measure (PROM) surveys are used to collect the data for calculating the numerator: 1) the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR) for THA patients, and 2) the Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR) for TKA patients.

These PROM data and specific risk variable data will be collected 90 to 0 days prior to surgery, and PROM data will be collected again 270 to 365 days following surgery.

Data elements used to define the numerator and for risk adjustment that are collected with PROM data include:

- HOOS, JR or KOOS, JR
- Date of Birth
- Single-Item Literacy Screening (SILS2) Questionnaire
- Body Mass Index (BMI) or Weight (kg) and Height (cm)
- Chronic (>90 Day) Narcotic Use
- Total Painful Joint Count (Patient-Reported in Non-Operative Lower Extremity Joint)
- Quantified Spinal Pain (Patient-Reported Back Pain, Oswestry Index Question)
- PROMIS Global Mental Health Score (calculated with data from the Patient-Reported

Outcomes Measurement Information Systems (PROMIS) Global or Veteran's Rand

12-Item Health Survey (VR-12); data from VR-12 is translated to PROMIS Global Mental

Health scores using a crosswalk created by Cella et. al for PROsetta® Stone)

(Please note: Data elements listed above are detailed in the Data Dictionary accompanying this NQF submission; see Tabs: Risk Variables with PRO Data; HOOS, JR; KOOS, JR; PROMIS Global; VR-12)

Center for Medicare and Medicaid Services (CMS) administrative data is used to identify eligible THA/

TKA procedures for the measure cohort (denominator) and additional risk variables, including patient demographics and clinical comorbidities (ICD-10 codes for eligible THA/TKA procedures identified in the Data Dictionary accompanying this NQF submission; see Tab ICD-10 2017-2018.)

The numerator is the risk-adjusted proportion of patients undergoing an elective primary THA/TKA that meet or exceed a SCB improvement on the HOOS, JR or KOOS, JR from preoperative to postoperative assessment. SCB improvement is defined as:

- For THA patients, an increase of 22 points or more on the HOOS, JR
- For TKA patients, an increase of 20 points or more on the KOOS, JR

SCB thresholds were defined using published literature (Lyman and Lee, 2018) and vetted by our Patient Working Group, TEP, and Technical Advisory Group.

Further, the measure numerator was defined with extensive patient and clinician input. Among the numerator definitions considered by stakeholders during measure development included:

- Change in PROM score from preoperative to postoperative assessment reported as an average for a hospital's patients;

- Postoperative PROM score reported as an average for a hospital's patients;

- A threshold change in PROM score from preoperative to postoperative assessment reported as a proportion of a hospital's patients meeting the threshold;

- A threshold postoperative PROM score reported as a proportion of a hospital's patients meeting the threshold; and

- A combination of both a minimum threshold change in PROM score from preoperative to postoperative assessment and a minimum threshold for postoperative PROM score.

Clinical experts and patients supported a numerator definition that assessed change in PROM score from preoperative to postoperative assessment over a numerator definition that focused on postoperative PROM score. TEP members and patients noted that patients want to see improvement, and that the numerator definition should reflect change following surgery. Comments against using a numerator definition focusing on the postoperative PROM score included concern that it does not reflect degree of improvement, and may incentivize surgery on patients with less severe disease who have better preoperative scores. This concern about assessment of the postoperative PROM score also led to dislike of the last option noted above, a numerator definition combining threshold change and threshold postoperative PROM score.

Stakeholders also strongly supported a numerator definition assessing a threshold change in PROM score over averaging patient change in PROM scores for hospital reporting. They noted that measurement of a threshold change will highlight lower performing patients, will protect at-risk patients, and is easy to understand as a performance measure. Comments against a reported average change included concern that a hospital whose patients all achieve average results could have a reported average change result that would be very similar to a hospital whose patients achieve either very good or very poor results; an average change numerator could show similar results for hospitals with very different patient outcomes).

The numerator definition of SCB threshold change, supported by patients and clinical experts, provides an easy to understand metric that patients found intuitive. Using a SCB threshold avoids the potential for misleading consumers and patients by averaging patient change scores across a hospital when individual patient outcomes within hospitals may vary considerably (as noted above). Using a SCB incentivizes providers to perform surgery on patients with worse baseline scores, a group that might otherwise not be offered surgery, as patients with poorer baseline PRO scores have more room to improve and thus a greater opportunity to achieve SCB. It also encourages providers to not perform THA/TKA procedures on patients with minimal symptoms, who will not benefit at all from surgery. And, since the SCB was defined with close input from patients and clinicians, it does set a minimum improvement threshold, but not one so large as to cause surgeons to avoid performing THA/TKA procedures on patients who would benefit.

References:

Cella D, Schalet BD, Kallen M, Lai JS, Cook KF, Rutsohn J, Choi SW. PROsetta® Stone Analysis Report Volume 2: A Rosetta Stone for Patient Reported Outcomes, PROMIS Global Health – Mental Component and VR-12 – Mental Component (Algorithmic Scores). http://www.prosettastone.org/LinkingTables1/GlobalHealth/Pages/default.aspx, 2018.

Lyman S and Lee YY. (2018). What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? Clin Orthop Relat Res, 467(12):2432-2441.

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

The cohort (target population) includes, Medicare fee-for-service (FFS) patients 65 years of age and older undergoing elective primary THA/TKA procedures, excluding patients with hip fractures, pelvic fractures and revision THAs/TKAs.

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).

The cohort for this measure is Medicare FFS patients 65 years of age and older undergoing an elective primary THA/TKA procedure at a non-federal short-term acute care hospital. Inclusion criteria includes patients:

- Enrolled in Medicare FFS Part A and Part B for the 12 months prior to the date of the index admission, and enrolled in Part A during the index admission
- Discharged alive from a non-federal short-term acute care hospital
- Undergoing only elective primary THA/TKA procedures (patients with fractures and revision procedures or with bone metastases are not included)
- Inclusion criteria are harmonized with CMS's existing measure cohort for the hospital-level 90-day risk-standardized THA/TKA complication measure

Center for Medicare and Medicaid Services (CMS) administrative data is used to identify qualifying THA/TKA procedures for the measure cohort. (ICD-10 codes for eligible THA/TKA procedures are identified in the Data Dictionary accompanying this NQF submission; see Tab ICD-10 2017-2018.)

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

Patients with staged procedures, defined as more than one elective primary THA or TKA performed on the same patient during distinct hospitalizations during the measurement period, are excluded. All THA/TKA procedures for patients with staged procedures during the measurement period are removed.

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.)*

Patients with staged procedures in the measure period are excluded. A staged procedure is identified if a patient has more than one hospitalization with an eligible, elective primary THA or TKA procedure during the measurement period. ICD-10 codes for eligible, elective primary THA/TKA procedures (listed in the Data Dictionary on "ICD-10 2017-2018" tab) are used to identify all eligible procedures during the measurement period; patients with an ICD-10 code for an eligible elective primary THA or TKA procedure in two or more hospital admissions during the measurement period are identified as having a staged procedure, and the patient, including all procedures, is removed from the measure cohort.

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 = Higher score

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

Target population: Medicare FFS patients 65 years and older undergoing an elective primary THA or TKA in a non-federal short-term acute care hospital.

To create the denominator:

Step 1. If the patient is a Medicare FFS patient, go to Step 2. If not, do not include in the denominator.

Step 2. If the patient is identified in CMS administrative claims data as having undergone an eligible elective primary THA or TKA during the measurement period, go to Step 3. If not, do not include in the denominator.

Step 3. If the patient is 65 years of age or older, go to Step 4. If not, do not include in the denominator.

Step 4. If the patient was enrolled in Medicare FFS Part A and Part B for the 12 months prior to index admission, and enrolled in Part A during the index admission, then go to Step 5. If not, do not include in the denominator.

Step 5. If the patient was discharged alive from the hospital, include in the denominator. If not, do not include in the denominator.

Step 6. If the patient experienced only one elective primary THA/TKA during the measurement period, or if the patient experience more than one elective primary THA/TKA during a singular hospitalization during the measurement period, + in the denominator. If the patient experienced two elective primary THA/TKA procedures during the measurement period performed during distinct hospitalizations, do not include in the denominator.

To create the numerator:

If the patient has complete PRO data collected during the prescribed preoperative and postoperative time windows, and meets or exceeds the SCB improvement threshold on the joint-specific PROM between the preoperative and postoperative assessment:

- for THA patients, an increase of 22 points on the HOOS, JR

- for TKA patients, an increase of 20 points on the KOOS, JR

then include in the numerator. If not, then do not include in the numerator.

The hospital-level measure result is calculated by aggregating all patient-level results across the hospital. For calculation of measure results, we recommend that hospitals should have a minimum case-volume of 25 elective primary THA/TKA patients with complete patient-reported outcomes and risk variable data collected 90 – 0 days preoperatively and complete patient-reported outcomes data collected 270 – 365 days postoperatively. Hospital-specific risk-standardized improvement rates (RSIRs) are calculated as the ratio of a hospital's "predicted" improvement to "expected" improvement multiplied by the overall observed improvement rate. Both predicted improvement and expected improvement are derived based on the output of a hierarchical logistic regression model that adjusts for patient case-mix and applies stabilized inverse probability weighting (IPW) to address potential non-response bias.

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.

This PRO-PM is not based on a sample. The measure will allow for proxy responses from a caregiver and hospitals will report whether the PROM survey responder is the patient or a surrogate.

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.

Collection of PRO data and accompanying risk variable data are to be collected 90 to 0 days prior to surgery, and postoperative PRO data are to be collected 270 to 365 days following surgery. The joint-specific PROM surveys (the HOOS, JR for THA patients and the KOOS, JR for TKA patients) are self-administered PRO surveys; some of the risk variable data are patient-reported (e.g., patient-reported back pain) and some are provider-reported (e.g., BMI). The preoperative collection window allows for data collection during preoperative visits while being near enough to the surgery to accurate reflect preoperative pain and functional status. The postoperative collection window allows for full recovery from THA or TKA surgery and aligns with postoperative physician visits for data collection. Whether PRO data are collected on paper surveys or electronically, data collection that aligns with physician office visits additionally allow for incorporation of PRO data into clinical care assessment and decision-making, increasing patient investment in data collection.

High response rates allow PRO-PMs to best represent hospital quality performance. Hospitals and physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Strong leadership support within the hospital, flexibility in rearranging clinical workflows to accommodate PRO data collection, accessibility of PRO data in real-time to inform clinical decision making can all increase staff investment in the value of PROs in improving care and quality, and PRO data used for clinical decisions can increase patient investment.

Hospital-level response rates for PRO data for this measure will be calculated as the percentage of elective primary THA or TKA procedures performed during the measurement period for which complete and matched preoperative and postoperative PRO and risk variable data have been submitted at each hospital; technically this is a submission rate, not a true response rate. A true response rate would consider how many patients were offered the opportunity to respond to the PRO survey and then, among those, how many actually responded. However, we are able to identify using claims data how many eligible patients undergo an elective primary THA/TKA during the measurement period and thus should have received a survey (defined by Medicare administrative claims data), excluding patients with staged procedures during the measurement period. Using this number as the denominator, we calculate an estimated response rate based on the number of complete PRO surveys (with complete clinician- and patient-reported risk variables) received.

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

If other, please describe in S.18.

Claims, Instrument-Based 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.

The PROM surveys used to define the measure outcome are 1) the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR) for THA patients, and 2) the Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR) for TKA patients. These instruments can be administered in paper or electronic form, filled out in person or over the phone. The HOOS, JR and KOOS, JR are presently available in English, not yet in other languages. For measurement of global mental health for risk adjustment, the Patient-Reported Outcomes Measurement Information System (PROMIS) Global or the Veterans RAND 12

Item Health Survey (VR-12) are used. The PROMIS Global is available in sixteen languages; the VR-12 is available in Spanish, Chinese and German.

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

Available in attached appendix at A.1

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

Facility

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

Inpatient/Hospital

If other:

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

N/A

2. Validity – See attached Measure Testing Submission Form

HipKneePROPMTestNQFForm_For_Submission.docx

2.1 For maintenance of endorsement

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

2.2 For maintenance of endorsement

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

2.3 For maintenance of endorsement

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

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

Measure Number (*if previously endorsed*)**:** TBD
Measure Title: Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)
Date of Submission: TBD

Type of Measure:

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

☐ Process *(including Appropriate Use)*	☐ Efficiency
☐ Structure	

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: Patient-reported survey data; Medicare Enrollment Database (EDB); Master Beneficiary Summary File (MBSF); American Community Survey data	☒ other: Patient-reported survey data; Medicare Enrollment Database (EDB); Master Beneficiary Summary File (MBSF); American Community Survey data

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 principal data for development and testing of this measure were patient-reported outcome (PROs) data and patient- and provider-reported risk variable data collected through the Center for Medicare and Medicaid Innovation (CMMI) Comprehensive Care for Joint Replacement (CJR) payment model. This model provided real-world data collection where participating hospitals received up to 2 points towards their overall Quality Score which was used to help determine model reconciliation payments. PRO data collection began in 2016 and calendar year (CY) 2020 will be the 5th and final performance year of the model. Dates for data collection noted in Section 1.3 (below) represent a combination of CJR model performance years 1 through 3 in order to capture a full 12 months of procedures with both preoperative and postoperative PRO data.

Additional data were used as follows:

Medicare Parts A and B claims data were used for identifying eligible elective primary THA/TKA procedures and for identifying patient comorbid conditions.

The Medicare Enrollment Database (EDB) was used to assess Medicare Fee for Service (FFS) enrollment and race. The Master Beneficiary Summary File (MBSF) was used to determine dual eligibility status. The Agency for Healthcare Research and Quality (AHRQ) socioeconomic status (SES) index score was derived from American Community Survey data.

Data from these data sources were linked for patients undergoing elective primary total hip arthroplasty (THA) and total knee arthroplasty (TKA) procedures from July 1, 2016 through June 30, 2017. Patients with complete preoperative and postoperative PRO and risk variable data were included in the dataset used for development and testing of this measure. These data were randomly divided 60%/40% into a Development Dataset and a Validation Dataset.

PRO data used for testing were collected consistent with PRO-PM measure specifications (PRO surveys, risk variable data elements and timing of preoperative and postoperative data collection were aligned).

1.3. What are the dates of the data used in testing? Click here to enter date range

This patient-reported outcome-based performance measure (PRO-PM) was tested on eligible procedures performed between July 1, 2016 and June 30, 2017. PRO and risk variable data were collected for patients 90 – 0 days prior to surgery and 270 – 365 days following surgery. Medicare claims between July 1, 2016 and June 30, 2017 were used to identify THA/TKA procedure codes, and Medicare claims for the 12 months prior to the procedure were used to identify a patient's comorbid conditions. The dates for Medicare Enrollment Database, Master Beneficiary Master File, and American Community Survey data to assess Medicare FFS status, socioeconomic indicators and race for patients were concurrent with their procedure data.

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, the measured entities are non-federal, acute inpatient US hospitals (including territories) serving Medicare FFS beneficiaries aged 65 years and older. Hospitals included in the development and testing of this PRO-PM were among those participating in CMMI's CJR payment model. A total of 238 hospitals submitted complete PRO and risk variable data for elective primary

THA/TKA procedures performed between July 1, 2016 and June 30, 2017 and these data were used for development and testing of this measure.

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 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 total dataset for measure development and testing included data from the 238 CJR participant hospitals that submitted complete preoperative and postoperative PRO and risk variable data for at least one elective primary THA/TKA procedure. Complete PRO and risk variable data was defined as the submission of preoperative patient-reported outcome measure (PROM) and risk variable data with no missing or out-of-range values for required data elements and that could be matched to postoperative PROM data with no missing or out-of-range values, for an elective primary THA/TKA procedure identified in claims data for the measurement period. The number of patients with complete PRO data for an elective primary THA or TKA procedure (excluding patients with staged elective primary THA/TKA procedures during the measurement period, defined as two or more procedures performed during separate inpatient admissions) was 11,270. These data were randomly divided 60%/40% into a Development Dataset and a Validation Dataset.

Development Dataset: This dataset includes 230 hospitals and 6,734 patients. Of these patients, 2,252 had a THA procedure and 4,482 had a TKA procedure. Characteristics of the 230 hospitals from which these data were submitted are presented in Table 1. Characteristics of the 6,734 patients in the dataset are presented in Table 2.

Validation Dataset: This dataset includes 219 hospitals and 4536 patients. Of these patients, 1,530 had a THA procedure and 3,006 had a TKA procedure. Characteristics of the 219 hospitals from which these data were submitted are presented in Table 1. Characteristics of the 4,536 patients in the dataset are presented in Table 2.

Combined Dataset (Hospitals ≥25 Patients with PRO Data): This dataset includes 123 hospitals from the total dataset with at least 25 THA/TKA patients with PRO data during the measurement period. (A case-volume cut-off of 25 was selected as it provided high measure result reliability and was consistent with volume thresholds used for public reporting of claims-based measures with which this measure was intentionally harmonized; we therefore recommend this measure be reported using a minimum case-volume cut-off of 25 or greater.) Table 1 identifies the characteristics of the 123 hospitals. These data were used for reliability and validity testing, and for response-bias analyses.

Table 1. Characteristics of Hospitals in Development and Validation Datasets and Hospitals with at least 25 THA/TKA Patients with PRO Data during the Measurement Period

Variable	Hospitals in Development Dataset	Hospitals in Validation Dataset	Hospitals in Combined Dataset for Reliability and Validity Testing and Response-Bias Analyses (with > 25 THA/TKA Patients with PRO Data)
Total Hospitals, N	230	219	123
Median # of Elective Primary THA/TKA Procedures Performed (Q1, Q3)	121 (56, 244)	123 (54, 250)	209 (114, 300)
Mean % of Patients on Medicaid (SD)	18.3%, 10.3	18.0%, 0.1	20.4%, 11.4
Region, %			
West	24.8%	25.1%	27.6%
Midwest	28.7%	31.1%	34.2%
Northeast	23.5%	21.9%	17.9%
South	23.0%	21.9%	20.3%
Teaching Status, %			
Teaching	46.1%	44.8%	48.8%
Non-Teaching	53.9%	55.2%	51.2%

Table 2. Patient Characteristics in Development and Validation Datasets

Variable	Development Dataset N (%)	Validation Dataset N (%)
Total N	6734	4536
Age in years *(Mean, SD)*	73.63 (5.75)	73.74 (5.84)
Sex: Male	2442 (35.97%)	1660 (36.60%)
Race: Black, non-Hispanic	254 (3.77%)	160 (3.53%)
White, non-Hispanic	6200 (92.07%)	4205 (92.70%)
Hispanic	178 (2.64%)	98 (2.16%)
Other	102 (1.51%)	73 (1.61%)
Bilateral procedure: Yes (vs. No)	31 (0.46%)	35 (0.77%)
Health Literacy (Comfort Filling Out Medical Forms by Yourself): None	1000 (14.85%)	663 (14.62%)
A little bit	518 (7.69%)	352 (7.76%)
Somewhat	775 (11.51%)	524 (11.55%)
Quite a bit	1192 (17.70%)	853 (18.81%)
Extremely	3249 (48.25%)	2144 (47.27%)
Patient-Reported Back Pain (Oswestry Index Question): None	2562 (38.05%)	1754 (38.67%)
Very Mild	1661 (24.67%)	1074 (23.68%)
Moderate	1706 (25.33%)	1156 (25.49%)
Fairly Severe	570 (8.46%)	391 (8.62%)

Variable	Development Dataset N (%)	Validation Dataset N (%)
Very Severe/Worst Imaginable	235 (3.49%)	161 (3.55%)
Patient-Reported Pain in Non-Operative Lower Extremity Joint: None	2298 (34.13%)	1552 (34.22%)
Mild	1640 (24.35%)	1125 (24.80%)
Moderate	1727 (25.65%)	1079 (23.79%)
Severe	856 (12.71%)	635 (14.00%)
Extreme	213 (3.16%)	145 (3.20%)
Body Mass Index (BMI) *(Mean, SD)*	30.39 (6.01)	30.46 (6.03)
Narcotic Use for >90 days	1224 (18.18%)	787 (17.35%)
PROMIS Global Mental Health Score *(Mean, SD)*	49.71 (8.10)	49.70 (8.05)
Severe Infection; other infectious diseases	1258 (18.68%)	842 (18.56%)
Diabetes or diabetes complications	1735 (25.76%)	1217 (26.83%)
Liver Disease	1794 (26.64%)	1229 (27.09%)
Rheumatoid arthritis and inflammatory connective tissue disease	750 (11.14%)	457 (10.07%)
Depression	1047 (15.55%)	698 (15.39%)
Other psychiatric disorders	1105 (16.41%)	714 (15.74%)
Coronary atherosclerosis or angina	1622 (24.09%)	1138 (25.09%)
Vascular or circulatory disease	1279 (18.99%)	862 (19.00%)
Renal failure	905 (13.44%)	621 (13.69%)

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.

Social Risk Factors available and analyzed included dual eligibility (dual Medicare and Medicaid coverage) and the AHRQ SES index.

Please note: We do not consider race a marker of socioeconomic status; we include it in our social risk factor analyses based upon literature specifically documenting racial and ethnic disparities in THA/TKA offer and acceptance rates as well as outcomes (Irgit and Nelson, 2011; Kerman et al, 2018).

Please also note: While health literacy also reflects social risk, our patient and technical experts strongly supported including health literacy in the risk model for a PRO-based measure, due to its very nature of asking patients to complete survey instruments as part of measurement. For this reason, we included it in the candidate risk variable list and in the final risk model; we therefore do not include health literacy in the specific social risk factor testing.

2a2. RELIABILITY TESTING

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

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)
☒ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)
☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

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

Data Element Reliability

Data element reliability is evidenced by reliability testing conducted during the development and validation of the joint-specific patient-reported outcome measures (PROMs) on which this THA/TKA PRO-PM is based.

HOOS, JR Reliability:

Internal consistency: The developers of the Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR) (Lyman et al, 2016a) assessed internal consistency reliability of using the Person Separation Index (PSI). The PSI was used in two data samples, the Hospital for Special Surgery (HSS) cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR), a nationally representative joint replacement registry. A higher value on the PSI indicates greater ability to differentiate patients with varying levels of ability, which in turn provides evidence of good internal consistency. For testing internal consistency for the HOOS, JR, a PSI value greater than 0.7 was considered acceptable (Lyman et al, 2016a). The developers also conducted principal component analysis on the standardized residuals to assess HOOS, JR items.

Test-retest reliability: Test-retest reliability was not tested by developers of the HOOS, JR as it had already been tested in the Hip dysfunction and Osteoarthritis Outcome Score (HOOS) in several validation studies (Klassbo et al, 2003; de Groot et al, 2007; Ornetti et al, 2010; Nilsdotter & Bremander, 2011). Intra-class correlation coefficients (ICCs) between dimensions (Pain, Symptoms, Activities of Daily Living, Sport and Recreation Function, and Quality of Life) were used to determine test-retest reproducibility.

KOOS, JR Reliability:

Internal consistency: The developers of the Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR) (Lyman et al, 2016b) assessed internal consistency reliability of using the PSI. The PSI was used in two data samples, the HSS cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR), a nationally representative joint replacement registry. A higher value on the PSI indicates greater ability to differentiate patients with varying levels of ability, which in turn provides evidence of good internal consistency. For testing internal consistency for the KOOS, JR, a PSI value greater than 0.7 was considered acceptable (Lyman et al, 2016b). The developers also conducted principal component analysis on the standardized residuals to assess KOOS, JR items.

Test-retest reliability: Test-retest reliability was not tested by developers of the KOOS, JR as it had already been tested in the Knee injury and Osteoarthritis Outcome Score (KOOS) (Roos et al, 1998). To examine test-retest reliability, the KOOS was administered to patients twice prior to surgery within a nine-day period. Intra-class correlation coefficients (ICCs) between dimensions (Pain, Symptoms, Activities of Daily Living, Sport and Recreation Function, and Quality of Life) were used to determine test-retest reproducibility.

Measure Score Reliability

The reliability of a measurement is the degree to which repeated measurements of the same entity agree with each other. For measures of hospital performance, the measured entity is the hospital, and reliability is the extent to which repeated measurements of the same hospital give similar results. Using the Combined Dataset (Development and Validation Datasets), we identified the hospitals with at least 25 THA/TKA patients with PRO data during the measurement period and assessed signal-to-noise reliability to describe how well the measure can distinguish performance of one hospital from another (Adams and Mehrota, 2010; Yu and Mehrota, 2013). The signal is the proportion of the variability in measured performance that can be explained by real differences in performance. 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.

References:

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

de Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SM, Favejee M, Roos EM, Verhaar JA. (2007). Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score. Osteoarthritis and Cartilage, 15:104-109.

Irgit, K., & Nelson, C. L. (2011). Defining Racial and Ethnic Disparities in THA and TKA. Clinical Orthopaedics and Related Research®, 469(7), 1817–1823.

Kerman, H. M., Smith, S. R., Smith, K. C., Collins, J. E., Suter, L. G., Katz, J. N., & Losina, E. (2018). Disparities in Total Knee Replacement: Population Losses in Quality-Adjusted Life-Years Due to Differential Offer, Acceptance, and Complication Rates for African Americans. Arthritis Care & Research, 70(9), 1326–1334.

Klässbo M, Larsson E, Mannevik E. (2003). Hip disability and osteoarthritis outcome score: An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scandinavian Journal of Rheumatology, 32(1), 46-51.

Lyman S, Lee YY, Franklin PD, Li W, Mayman DJ, Padgett DE. (2016a). Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1472-1482.

Lyman S, Lee YY, Franklin PD, Li W, Cross MB, Padgett DE. (2016b). Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1461-1471.

Nilsdotter A, Bremander A. (2011). Measures of hips function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity of Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. Arthritis Care & Research, 63(S11): S200-S207.

Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, Guillemin F, Maillefert JF. (2010). Cross-cultural adaptation and validation of the French version of the Hip disability and Osteoarthritis Outcome Score (HOOS) in hip osteoarthritis patients. Osteoarthritis and Cartilage, 18:522-529.

Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. (1998). Knee Injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. J Orthop Sports Phys Ther, 8(2):88-96.

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

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*)

<u>Data Element Reliability Results</u>

Data element reliability results are reported for reliability testing conducted during the development and testing of the joint-specific PROMs on which this THA/TKA PRO-PM is based.

HOOS, JR Reliability:

Internal consistency: The developers of the HOOS, JR (Lyman et al, 2016a) assessed internal consistency reliability of using the Person Separation Index (PSI). Internal consistency of the HOOS, JR on the PSI were 0.86 in the HSS cohort and 0.87 in the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) cohort. Results of a principal component analysis conducted on the standardized residuals indicated that the six HOOS, JR items existed in a single dimension (Lyman et al, 2016a).

Test-retest reliability: Test-retest reliability was not tested by developers of the HOOS, JR as it had already been tested in the Hip dysfunction and Osteoarthritis Outcome Score (HOOS) in several validation studies (Klassbo et al, 2003; de Groot et al, 2007; Ornetti et al, 2010; Nilsdotter & Bremander, 2011). Intra-class correlation coefficients (ICCs) were used to determine test-retest reproducibility and ranged from 0.75 to 0.97 in the validation studies. Specifically, the Pain and Activity of Daily Living domains, from which HOOS, JR pain and functioning questions are drawn, had ICCs of 0.83 - 0.89 (Pain sub-scale) and 0.86 - 0.94 (Activity of Daily Living sub-scale).

KOOS, JR Reliability:

Internal consistency: The developers of the KOOS, JR (Lyman et al, 2016b) assessed internal consistency reliability of using the PSI. Internal consistency of the KOOS, JR on the PSI were 0.84 in the HSS cohort and 0.85 in the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) cohort. Results of a principal component analysis conducted on the standardized residuals indicated that the seven KOOS, JR items existed in a single dimension (Lyman et al, 2016b).

Test-retest reliability: Test-retest reliability was not tested by developers of the KOOS, JR as it had already been tested in the Knee injury and Osteoarthritis Outcome Score (KOOS) (Roos et al, 1998). Intra-class correlation coefficients (ICCs) were used to determine test-retest reproducibility and ranged from 0.75 to 0.93. Specifically, the Pain, Activity of Daily Living and Symptom domains, from

which KOOS, JR pain, functioning and stiffness questions are drawn, had ICCs of 0.85 (Pain sub-scale), 0.75 (Activity of Daily Living sub-scale), and 0.93 (Symptoms).

Measure Score Reliability Results

The signal-to-noise ratio (see Table 3, below) yielded a median reliability score of 0.9589 (range: 0.8956 – 0.9916). Interquartile range was 0.0370.

References:

de Groot IB, Reijman M, Terwee CB, Bierma-Zeinstra SM, Favejee M, Roos EM, Verhaar JA. (2007). Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score. Osteoarthritis and Cartilage, 15:104-109.

Klässbo M, Larsson E, Mannevik E. (2003). Hip disability and osteoarthritis outcome score: An extension of the Western Ontario and McMaster Universities Osteoarthritis Index. Scandinavian Journal of Rheumatology, 32(1), 46-51.

Lyman S, Lee YY, Franklin PD, Li W, Mayman DJ, Padgett DE. (2016a). Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1472-1482.

Lyman S, Lee YY, Franklin PD, Li W, Cross MB, Padgett DE. (2016b). Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1461-1471.

Nilsdotter A, Bremander A. (2011). Measures of hips function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity of Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire. Arthritis Care & Research, 63(S11): S200-S207.

Ornetti P, Parratte S, Gossec L, Tavernier C, Argenson JN, Roos EM, Guillemin F, Maillefert JF. (2010). Cross-cultural adaptation and validation of the French version of the Hip disability and Osteoarthritis Outcome Score (HOOS) in hip osteoarthritis patients. Osteoarthritis and Cartilage, 18:522-529.

Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. (1998). Knee Injury and Osteoarthritis Outcome Score (KOOS)—development of a self-administered outcome measure. J Orthop Sports Phys Ther, 8(2):88-96.

Table 3. Signal to Noise Reliability, Hospitals with Volume ≥25 THA/TKA Patients with PRO Data

Reliability Statistic	# Hospitals	Median	Mean (SD)	Min	Max	Interquartile Range		
						Q1	Q3	Range
Signal-to-noise	123	0.9589	0.9520 (0.263)	0.8956	0.9916	0.9351	0.9717	0.0366

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?*)

Data Element Reliability

The reliability results from the literature demonstrate that the HOOS, JR and the KOOS, JR PROM instruments are sufficiently reliable and exceed accepted norms for reliability testing. The results

assessing internal consistency indicated PSI values of 0.86 - 0.87 for the HOOS, JR (Lyman et al, 2016a) and 0.84-0.85 for the KOOS, JR, (Lyman et al, 2016b) indicating values well above 0.7, indicating the ability of the instruments to differentiate patients with varying levels of pain and functioning, which in turn provides evidence of good internal consistency. Test-retest reliability results for the HOOS domains from which HOOS, JR questions were drawn (Pain and Activity of Daily Living domains) revealed high intra-class correlation coefficients (ICCs). Likewise, test-retest reliability for the KOOS domains from which the KOOS, JR questions were drawn (ICCs of 0.75 - 0.93) provided evidence good reliability.

Measure Score Reliability

The signal-to-noise reliability of 0.96 indicates excellent reliability.

Our interpretation of these results is based on standards established by Landis and Koch (1997):

<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
1 = Perfect agreement

References:

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

Lyman S, Lee YY, Franklin PD, Li W, Mayman DJ, Padgett DE. (2016a). Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1472-1482.

Lyman S, Lee YY, Franklin PD, Li W, Cross MB, Padgett DE. (2016b). Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®*, 474(6):1461-1471.

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)*

Data Element Validity

Data element validity is evidenced by validity testing conducted during the development and testing of the joint-specific PROMs on which this THA/TKA PRO-PM is based. All validity testing for the HOOS, JR and KOOS, JR instruments was conducted by the PROM developers (Lyman et al, 2016a; Lyman et al, 2016b).

HOOS, JR Validity:

Responsiveness: Responsiveness of the HOOS, JR to changes following a total hip replacement was evaluated using standardized response means, and then examined against other previously validated PROMs (HOOS domains, The Western Ontario and McMaster University Arthritis Index [WOMAC] domains) in the HSS cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR) registry at 2 years after a THA procedure (Lyman et al, 2016a). A standardized response mean greater than 0.8 was considered large (Steiner and Norman, 2003).

External validity: External construct validity was evaluated using Spearman's correlations between HOOS, JR and the HOOS and the WOMAC. A Spearman's correlation coefficient of 0.8 or greater was considered very high external validity (Wechsler, 1996). External correlations were assessed using a scatterplot overlying a contour plot based on bivariate kernel density estimation between the HOOS, JR and HOOS domains (Lyman et al, 2016a).

Floor and ceiling effects: Floor and ceiling effects (percent at worst possible score preoperatively and best possible score postoperatively) were evaluated against the HOOS and the WOMAC instruments (Lyman et al, 2016a).

KOOS, JR Validity:

Responsiveness: Responsiveness of the KOOS, JR to changes following total knee replacement was evaluated using standardized response means, and then examined against other validated PROMs (KOOS domains, WOMAC domains) in the validation cohort (Lyman et al, 2016b). A standardized response mean greater than 0.8 was considered large (Steiner and Norman, 2003).

External validity: External construct validity was evaluated using Spearman's correlations between KOOS, JR and the KOOS and the WOMAC. A Spearman's correlation coefficient of 0.8 or greater was considered very high external validity (Wechsler, 1996). External correlations were assessed using a scatterplot overlying a contour plot based on bivariate kernel density estimation between the KOOS, JR and KOOS domains (Lyman et al, 2016b).

Floor and ceiling effects: Floor and ceiling effects (percent at worst possible score preoperatively and best possible score postoperatively) were evaluated against the KOOS and the WOMAC instruments (Lyman et al, 2016b).

Empirical Measure Score Validity

To assess empirical measure score validity we compared the THA/TKA PRO-PM risk-standardized improvement rates (RSIRs) to the NQF endorsed Hip/Knee Complication Measure (NQF #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA.) This measure estimates the risk-adjusted rate that patients who have experienced an elective primary THA/TKA experience at least one of eight complications within 90 days of the procedure. The RSCR is categorized into 3 groups: worse than national average, same as national average, and better than national average. Data for the hospital RSCRs from April 1, 2015 to March 31, 2018 were compared to RSIRs for procedures performed July 1, 2016 to June 30, 2017.

References:

Lyman S, Lee Y-Y, Franklin PD, Li W, Mayman DJ, Padgett DE. Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1472-1482.

Lyman S, Lee Y-Y, Franklin PD, Li W, Cross MB, Padgett DE. Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1461-1471.

Steiner DL, Norman GR. (2003). Health Measurement Scales: A Practical Guide to Their Development and Use. London, UK: Oxford University Press.

Wechsler S. (1996). Statistics at Square One. 9th ed. London, UK: BMJ Publishing Group.

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

Data Element Validity

Data element validity results are reported for validity testing conducted during the development and testing of the joint-specific PROMs on which this THA/TKA PRO-PM is based.

HOOS, JR Validity:

Responsiveness: Standardized response means for the HOOS, JR relative to other PROMs measuring post-surgery hip improvement were 2.38 (95% CI, 2.27–2.49) in the HSS data and 2.03 (95% CI, 1.84–2.22) in the FORCE registry data.

External validity: Correlations between the HOOS, JR and HOOS Pain domain were 0.87 (95% CI, 0.86–0.89) in the HSS data and 0.87 (95% CI, 0.84–0.90) in the FORCE registry data. Correlations between the HOOS, JR and HOOS Activity of Daily Living domain were 0.94 (95% CI, 0.93–0.95) in the HSS data and 0.94 (95% CI, 0.93–0.96) in the FORCE registry data. Likewise, correlations between the HOOS, JR and the WOMAC Pain domain was 0.84 (95% CI, 0.81–0.86) in the HSS data and 0.85 (95% CI, 0.81–0.88) in the FORCE registry data; between HOOS, JR and WOMAC Functioning were 0.94 (95% CI, 0.93–0.95) in the HSS data and 0.94 (95% CI, 0.93–0.96) in the FORCE registry data; and between the HOOS, JR and WOMAC Stiffness domain were 0.64 (95% CI, 0.58–0.71) in the HSS data and 0.65 (95% CI, 0.61–0.68) in the FORCE registry data (Lyman et al, 2016a).

Floor and ceiling effects: The HOOS, JR showed floor (0.6%–1.9%) and ceiling (37%–46%) effects, and were comparable to or better than HOOS domains and the WOMAC (Lyman et al, 2016a).

KOOS, JR Validity:

Responsiveness: Standardized response means for the KOOS, JR relative to other PROMs measuring post-surgery knee improvement were 1.79 (95% CI, 1.70–1.88) in the HSS data and 1.70 (95% CI, 1.54–1.86) in the FORCE registry data.

External validity: Correlations between the KOOS, JR and KOOS Pain domain were 0.89 (95% CI, 0.88–0.91) in the HSS data and 0.91 (95% CI, 0.90–0.93) in the FORCE registry data. Correlations between the KOOS, JR and KOOS Activity for Daily Living domain were 0.87 (95% CI, 0.85–0.88) in the HSS data and 0.84 (95% CI, 0.81–0.87) in the FORCE registry data. Correlations with the Symptoms domain were 0.59 (95% CI, 0.55–0.64) in the HSS data and 0.69 (95% CI, 0.64–0.74) in the FORCE registry data. Similarly, correlations between the KOOS, JR and WOMAC Pain were 0.80 (95% CI, 0.77–0.82) in the HSS data and 0.82 (95% CI, 0.79–0.86) in the FORCE registry data; between KOOS, JR and WOMAC Function were 0.87 (95% CI, 0.85-0.88) in the HSS data and 0.84 (95% CI, 0.81–0.87 in the FORCE registry data; and between KOOS, JR and WOMAC Stiffness were 0.72 (95% CI, 0.69–0.75 in the HSS data and 0.76 (95% CI, 0.72–0.80) in the FORCE registry data (Lyman et al, 2016b).

Floor and ceiling effects: Floor effects for the KOOS, JR (percent at worst possible score preoperatively) were 0.4 – 1.2% and the ceiling effects (percent at best possible score postoperatively) were 18.8 – 21.8% (Lyman et al, 2016b).

<u>Empirical Measure Score Validity</u>

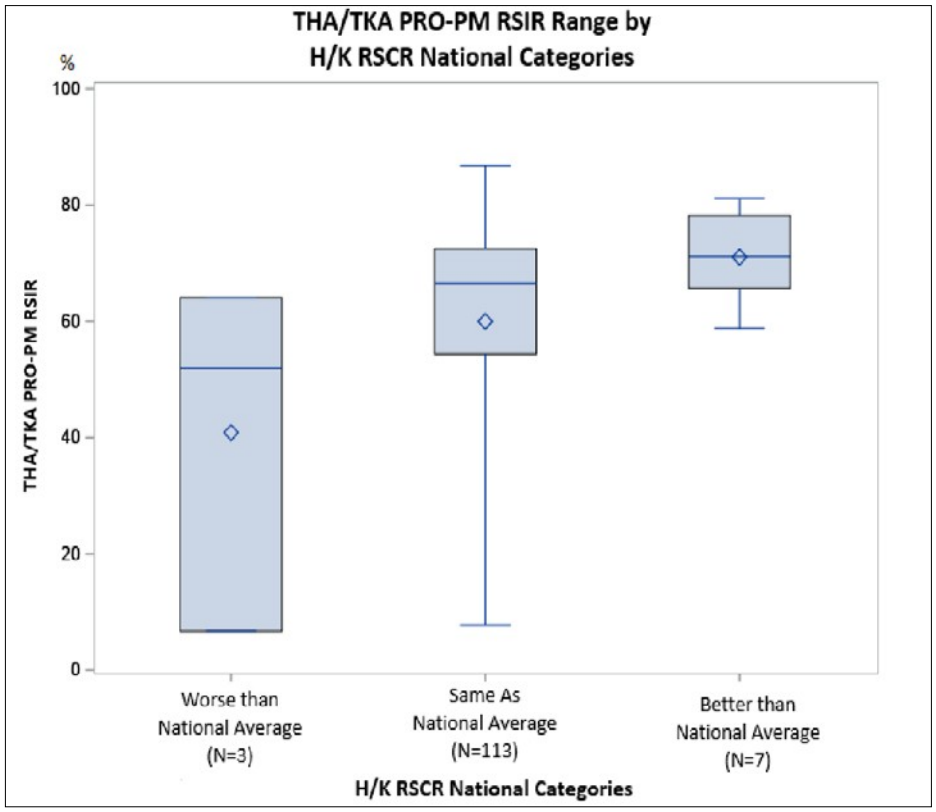
Comparison of THA/TKA PRO-PM RSIRs to RSCR categories indicates an increasing monotonic trend. Those hospitals in the "RSCR Worse than National Average" category have lower median RSIRs (51.87%) than the median RSIR (66.49%) of hospitals in the "RSCR Same as National Average" category, which is lower than that of hospitals in the "RSCR Better than National Average" category (71.13%). The hospitals with lower risk-adjusted complication rates have higher risk-adjusted THA/TKA improvement rates (Figure 1).

<u>References:</u>

Lyman S, Lee Y-Y, Franklin PD, Li W, Mayman DJ, Padgett DE. Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1472-1482.

Lyman S, Lee Y-Y, Franklin PD, Li W, Cross MB, Padgett DE. Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1461-1471.

Figure 1. Range of THA/TKA PRO-PM RSIRs by THA/TKA RSCR National Categories within Hospitals Submitting Complete PRO Data (Hospitals with ≥25 THA/TKA Procedures, N=123)



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?*)

Data Element Validity

The validity results from the literature demonstrate that the HOOS, JR and the KOOS, JR PROM instruments are valid and meaningful measures for assessing patient-reported outcomes following THA/TKA procedures. The HOOS, JR and the KOOS, JR showed very high responsiveness, well beyond the 0.8 standardized response mean value considered "very large" (Steiner and Norman, 2003). Spearman correlation values between the HOOS, JR and the HOOS domains from which the HOOS, JR questions were drawn (Pain and Activity of Daily Living domains) were high; likewise, Spearman correlation values between the KOOS, JR and the KOOS Pain and Activity of Daily Living domains were high, and were moderate between the KOOS, JR and the Symptom domain. Floor effects were small; ceiling effects for the HOOS, JR were 37%–46%, but were comparable to or better than HOOS domains and the WOMAC (Lyman et al, 2016a; Lyman et al, 2016b).

Empirical Measure Score Validity

As these outcomes are not clinically expected to be perfectly correlated but do reflect hospital-level care and processes impacting quality of care for patients experiencing elective primary THA/TKA surgery, we interpret the increasing monotonic trend between RSIRs and RSCR national categories as reflective of empiric measure validity. As NQF is aware, empiric validation of novel outcome

measures is challenging as there is rarely, if ever, a "gold standard" against which to compare the measure.

References:

Lyman S, Lee Y-Y, Franklin PD, Li W, Mayman DJ, Padgett DE. Validation of the HOOS, JR: A Short-form Hip Replacement Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1472-1482.

Lyman S, Lee Y-Y, Franklin PD, Li W, Cross MB, Padgett DE. Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey. *Clinical Orthopaedics and Related Research®.* 2016;474(6):1461-1471.

Steiner DL, Norman GR. (2003). Health Measurement Scales: A Practical Guide to Their Development and Use. London, UK: Oxford University Press.

2b2. EXCLUSIONS ANALYSIS

NA ☐ no exclusions — *skip to section 2b3*

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*)

Patients with staged procedures, defined as two or more elective primary THA or TKA procedures performed on the same patient during distinct hospitalizations during the measurement period, were excluded from the measure. The overlapping recovery periods for staged procedures occurring within one year of each other has two consequences that set patients experiencing staged THA/TKA procedures apart from patients experiencing unilateral or bilateral procedures: 1) the recovery from one procedure may negatively impact recovery from the other procedure; and 2) it may be challenging to fully distinguish the recovery for either of the procedures from the other with postoperative PRO data (collected 270 to 365 days after surgery). For these reasons, patients with staged procedures during the measurement period were excluded from the denominator.

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*)

Among the 238 hospitals in the total dataset, 491 (4.17%) of patients with complete PRO and risk variable data for staged procedures during the measurement period were excluded. Across hospitals, the mean proportion of procedures excluded from the analysis was 3.84% (SD 5.69), and the median proportion was 2.11%.

Table 4. Proportion of Procedures Across Hospitals Removed for Exclusion of Staged Procedures (Total Dataset)

Summary Statistics	Staged Procedures
N (Hospitals)	238
N (THA/TKA Patients with Complete PRO and Risk Variable Data)	11761
N (Staged Procedures with Complete PRO and Risk Variable Data)	491
Mean (SD)	3.84% (5.69)
Percentile	
100% Max	43.48%
99%	33.33%
75% (Q3)	5.26%
50% (Median)	2.11%
25% (Q1)	0.00%
1%	0.00%
0% Min	0.00%

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*)

The exclusion of staged procedures from the analysis removes a potential negative impact on hospital-specific measure results since the recovery from one procedure may negatively impact recovery from the other procedure, and the challenge of fully distinguishing the recovery for either of the procedures from the other with postoperative PRO data (collected 270 to 365 days after surgery). While bilateral procedures share the same follow-up period and can be accounted for in the risk model (and thus are not excluded), staged procedures that are performed at distinct times with varying amounts of time between procedures per patient make accurate risk adjustment challenging.

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 2b4.

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

☐ **No risk adjustment or stratification**
☒ **Statistical risk model with 19 risk factors**
☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

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.

For model development we used a logistic regression model, with outcome Y_i for the i^{th} patient equal to 1 if the patient had achieved substantial clinical benefit (SCB) improvement on the PROM score from preoperative to postoperative assessment, and zero otherwise. [Substantial clinical benefit improvement is measured as a 22-point increase on the HOOS, JR from preoperative to postoperative assessment for THA patients, and a 20-point increase on the KOOS, JR from preoperative to postoperative assessment for TKA patients.] We developed this model using risk variables identified in a systematic literature review/environmental scan and by orthopedists surveyed about what risk variables they consider important in predicting THA/TKA outcomes that were then prioritized by our technical expert panel (TEP) and clinical experts as both clinically important and feasible.

Please note: A table of initial candidate risk variables under consideration and those finalized in the CJR Final Rule in 2015 (Centers for Medicare & Medicaid Services, CJR Final Rule 2015) for PRO data collection efforts for this measure are included in the Data Dictionary accompanying this NQF submission. This table (Table 4) includes proposed data sources for each variable, and ratings for each on five selection criteria: availability, importance, ease of collection, reliability, and validity. This list of variables was reviewed and prioritized by the TEP and other stakeholders and clinical experts. A second table, also included in the Data Dictionary (Table 5), identifies the list of risk variables available from CJR PRO data collection and from Medicare claims data assessed in risk modeling analyses during measure development. [Further details about risk variable selection for the risk model described in Section 2b3.3a, below.]

The risk variables included in the final model are:

- **Age, in Years**
- **Male Sex**
- **Procedure: THA**
- **Bilateral procedure**
- **Health Literacy (assessed by response to Single Item Literacy Screener questionnaire, "Comfort Filling Out Medical Forms by Yourself") (Wallace et al, 2006; Sarkar et al, 2011)**
- **Back Pain at preoperative assessment (Quantified Spinal Pain: Patient-Reported Back Pain, Oswestry Disability Index question) (Fairbank et al, 2000; Ayers et al, 2013)**
- **Pain in Non-Operative Lower Extremity Joint (Total painful joint count: Patient-Reported in Non-operative Lower Extremity Joint) (Ayers et al, 2013)**
- **Body Mass Index, in kg per m^2**
- **Narcotic Use for >90 days**
- **Baseline PROMIS Global Mental Health Subscale Score**
- **Severe infection; other infectious diseases (CC 1, 3-7)**
- **Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)**
- **Liver disease (CC 27-31)**
- **Rheumatoid arthritis and inflammatory connective tissue disease (CC 40)**
- **Depression (CC 61)**
- **Other psychiatric disorders (CC 63)**
- **Coronary atherosclerosis or angina (CC 88-89)**
- **Vascular or circulatory disease (CC 106-109)**
- **Renal failure (CC 135-140)**

We estimated the hospital-specific RSIR using a hierarchical logistic regression model (hierarchical model). This strategy accounts for within-hospital correlation of the observed outcome among patients and accommodates the assumption that underlying differences in the quality of care across hospitals lead to systematic differences in patient outcomes. This approach models the log

odds of patient improvement on the PROM as a function of patient demographics and clinically relevant comorbidities with an intercept for the hospital-specific random effect.

We then calculate the hospital-specific RSIRs, which were calculated as the ratio of a hospital's "predicted" number of improvements to "expected" number of improvements multiplied by the overall observed improvement rate. The expected number of improvements for each hospital (denominator) was estimated using its patient mix and the average hospital-specific intercept (the average intercept among all hospitals in the sample). The predicted number of improvements for each hospital (numerator) was estimated given the same patient mix but an estimated hospital-specific intercept. Operationally, the expected number of improvements for each hospital was obtained by summing the expected improvement for all patients in the hospital. The expected improvement for each patient was calculated via the hierarchical model, which applies the estimated regression coefficients to the observed patient characteristics and adds the average of the hospital-specific intercept. The predicted number of improvement for each hospital was calculated by summing the predicted improvement for all patients in the hospital. The predicted improvement for each patient was calculated through the hierarchical model, which applies the estimated regression coefficients to the patient characteristics observed and adds the hospital-specific intercept.

More specifically, we used a hierarchical logistic regression model to account for the natural clustering of observations within hospitals. The model employs a logit link function to link the risk factors to the outcome with a hospital-specific random effect:

Let YY_{iiii} denote the outcome (equal to 1 if patient has an improvement, zero otherwise) for patient *i* at hospital *j*; ZZ_{iiii} denotes a set of risk factors for patient ii at hospital ii, nn_{ii} is the number of index admissions to hospital ii. We assume the outcome is related linearly to the covariates via a logit function:

Logistic Regression Model

$$\text{logit}\{PPPPPPPP(YY_{iiii}=11)\}=\alpha\alpha+\beta\beta ZZ_{iiii} \qquad \textbf{(1)}$$

and $ZZ_{iiii}=(ZZ_{11iiii}, ZZ_{22iiii}, \ldots, ZZ_{ppiiii})$ is a set of pp patient-specific covariates.

To account for the natural clustering of observations within hospitals, we estimate a hierarchical logistic regression model that links the risk factors to the same outcomes and a hospital-specific random effect.

Hierarchical Logistic Regression Model

$$llllllliiil\{PPPPPPPP(YY_{iiii}=11)\}=\alpha\alpha_{ii}+\beta\beta ZZ_{iiii} \qquad \textbf{(2)}$$

$$\textbf{where } \alpha\alpha_{ii}=\mu\mu+\omega\omega_{ii};\ \omega\omega_{ii}\sim NN(00,\tau\tau^{22}) \qquad \textbf{(3)}$$

where $\alpha\alpha_{ii}$ represents the hospital-specific intercept, ZZ_{iiii} is defined as above, μ is the adjusted average intercept over all hospitals in the sample, $\omega\omega_{ii}$ is the hospital-specific intercept deviation from $\mu\mu$, and τ^2 is the between-hospital variance component. This model separates within-hospital variation from between-hospital variation. Both the hierarchical logistic regression model and logistic regression model are estimated using the SAS software system (GLIMMIX and LOGISTIC procedures, respectfully).

We first fit the logistic regression model described in Equation (1) in selecting covariates in the best model. Having identified the covariates that remained, we then apply stabilized inverse probability weights (IPW) that are calculated from a propensity score analysis using multinomial logistic regression to model three PRO data response groups: complete PRO submission, incomplete PRO submission, and no response. (See 2b6.1 for a detailed description of the analytic approach to addressing potential response bias.) Next, we fit the hierarchical logistic regression model described in Equations (2) and (3) to the corresponding parameters. Lastly, we calculate the risk-standardized improvement rate in the way described above.

Thus, at the hospital level, this measure will be calculated and presented as a RSIR, producing a performance measure per hospital which accounts for patient case-mix and applies stabilized inverse probability weighting (IPW) to address potential non-response bias and represents a measure of quality of care following primary elective THA and TKA. Response rates for PRO data for this measure will be calculated as the percentage of elective primary THA or TKA procedures for which complete and matched preoperative and postoperative PRO data have been submitted divided by the total number of eligible THA or TKA procedures performed at each hospital.

References:

Ayers DC, Li W, Oatis C, Rosal MC, Franklin PD. (2013). Patient-reported outcomes after total knee replacement vary on the basis of preoperative coexisting disease in the lumbar spine and other nonoperatively treated joints: the need for a musculoskeletal comorbidity index. *The Journal of bone and joint surgery American volume,* 95(20):1833.

Comprehensive Care for Joint Replacement (CJR) Payment Model for Acute Care Hospitals Furnishing Lower Extremity Joint Replacement Services Final Rule, 80 C.F.R. 73273 (Nov 24, 2015).

Fairbank JCP, Paul B. (2000). The Oswestry disability index. *Spine,*25(22):2940-2953.

Sarkar U, Schillinger D, López A, Sudore R. Validation of self-reported health literacy questions among diverse English and Spanish-speaking populations. J Gen Intern Med. 2011 Mar;26(3):265-71. Epub 2010 Nov 6.

Wallace LS, Rogers ES, Roskos SE, Holiday DB, Weiss BD. Brief report: screening items to identify patients with limited health literacy skills. J Gen Intern Med. 2006 Aug;21(8):874-7.

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

N/A

2b3.3a. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors (clinical factors or social risk factors) used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care*) **Also discuss any "ordering" of risk factor inclusion**; for example, are social risk factors added after all clinical factors?

We identified risk variables from the published literature through a systematic literature review and environmental scan, as well as from orthopedists surveyed about what risk variables they consider important in predicting THA/TKA outcomes and their feasibility based on common clinical practice. In consultation with the Technical Working Group and the TEP and through detailed public comments from specialty societies, we focused on candidate risk-adjustment variables of interest that were clinically relevant, reliably and standardly collected in clinical care, and had an evidence-based relationship with clinical outcomes following elective primary THA or TKA.

We used the comprehensive list of candidate risk variables obtained through expert and public input to survey our TEP on their thoughts to each risk variable's priority. In addition, we collaborated with orthopedic societies and individual orthopedic practices to evaluate the feasibility, uniformity and reliability of clinical data elements prioritized by orthopedists by performing a medical record review at seven practices across the country.

In addition to clinical risk variables that have been collected de novo and evaluated for inclusion in the final measure risk model, all diagnostic codes from administrative claims during the 12 months prior to the THA/TKA procedure were evaluated for possible inclusion in the risk model.

The burden of novel data collection for PRO-based performance measures adds complexity to risk adjustment for this measure as the measure will also need to account for non-response and/or incomplete data and the overall response rate at each hospital. We recognize that poorly or incompletely collected data may be asymmetrically distributed across lower socioeconomic or disadvantaged populations with the potential to directly affect measure scores. Although sociodemographic factors also potentially affect other outcome measures, PRO-based measures are particularly vulnerable to these factors, most specifically health literacy.

The principles underlying the assessment of individual risk variables in the context of risk model development are summarized below:

- **The goal of risk adjustment is to account for patient characteristics that are reasonably beyond the control of the hospital. Therefore, risk variables must represent clinically important risk predictors; that is, they must be predictive of the outcome (in this case, the change in PROs after THA/TKA) and reasonably beyond hospital control.**
    - **The goal is not perfect risk prediction – this would imply that the hospital has no impact on clinical outcomes (that is, all variation is entirely explained by patient characteristics and healthcare providers have no impact on clinical outcomes). We know this is not true – providers can improve care and outcomes through active quality improvement efforts (such as patient education, adjustments to patient care before, during and after surgery).**

- **Risk variables must be feasible to collect and report. If a variable creates a data collection burden to patients, surgeons, hospitals, or the healthcare system, the incremental value of including the variable in the risk model should significantly outweigh the burden.**
    - **The definition of burden is subjective. This measure can only be implemented by requiring that hospitals, surgeons, and patients collect the PROM and relevant risk variables data both before and after the THA/TKA. The TEP recommended that we collect both a global PROM (the PROMIS Global or VR-12) and a hip- or knee-specific PROM (the HOOS, JR or KOOS, JR). It is our goal to minimize any *additional* data collection requirements beyond the PROM surveys, if possible.**
- **Risk variables must be reliably and consistently defined so that the risk variables carry the same information across all patients and hospitals.**

Finally, we will only include risk variables that have been tested empirically in the preliminary risk model. If risk factors are important but unavailable, we can either test available surrogate risk factors and/or CMS can pursue additional data collection for future iterations of the measure. Through our extensive stakeholder engagement that informed prospective data collection through CJR, we believe we have access to sufficiently exhaustive risk variable data to inform a robust risk model.

To select the final risk model, we surveyed the TEP and asked them to rank the importance of clinical variables for use in a PRO-PM risk model. We solicited additional input from clinical consultants to create a list of clinically relevant and important risk variables for risk adjustment of a THA/TKA PRO-PM. We assessed model performance in the Development Dataset examining the model performance (C-statistics), model calibration (lack of fit), model discrimination in terms of predictive ability (range of observed outcome among deciles of predicted outcomes), and distribution of model residuals. We calculated the model estimates as well as the coefficients and 95% confidence intervals for risk-adjustment variables for the best-performing model in the development dataset. We assessed risk factors in THA-specific and TKA-specific cohorts to ensure risk prediction for a combined THA/TKA cohort was consistent with that for THA- and TKA-specific cohorts. We compared measure results and risk model performance for the THA- and TKA-specific and the combined THA/TKA cohorts. We then repeated assessment of model performance for the final combined THA/TKA cohort in the Validation Dataset.

To address non-response bias, we identified variables associated with non-response to PRO survey data in two ways. First, we identified statistical associations of patient characteristics and clinical comorbidities in our data across three PRO response groups: patients with complete PRO data submission, patients with incomplete PRO data submission, and patients with no response. Next, we conducted a literature review and identified variables associated with unit non-response to PROM survey data by other investigators, selecting to include variables identified in the literature that were likewise available in our data. (See 2b6.1 for a detailed description of the analytic approach to addressing potential response bias.)

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?

Testing results using the Validation Dataset of the final risk-adjusted model for SCB improvement following elective primary THA/TKA are presented in Table 5, below. Risk variable odds ratios (ORs) are adjusted for other risk variables in the model and are adjusted with non-response weighting to address response bias. As previously noted, the SCB outcome allows patients with poor baseline PRO scores to improve, so some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB.

Testing comparing THA- and TKA-specific risk model results to the final combined THA/TKA risk model (which includes all risk variables included in the THA- and TKA-specific risk models) demonstrated that model performance was equal or better in the combined THA/TKA cohort (Table 6, below). Pearson's correlation coefficient for hospital-level RSIRs calculated with the THA-specific risk model compared to RSIRs calculated with the combined THA/TKA risk model was excellent at 0.945 (p<0.001) (see Figure 2, below). Likewise, Pearson's correlation coefficient for RSIRs calculated with the TKA-specific risk model compared to those for the combined THA/TKA risk model was excellent at 0.976 (p<0.001) (see Figure 3, below).

Table 5. Final Risk Model Variables and Adjusted Odds Ratios (HLM): Validation Dataset (Patient N = 4,536, Hospital N = 219)

Variable	Frequency	OR (95% CI) (Weighted for Non-Response)
Age	*Mean*=73.74 (*SD* 5.84)	1.00 (0.99, 1.01)
Sex: Male	1660 (36.60%)	**0.78 (0.67, 0.90)**
Procedure: THA	1530 (33.73%)	**1.40 (1.19, 1.64)**
Bilateral procedure	35 (0.77%)	1.42 (0.84, 2.40)
Health Literacy (Comfort Filling Out Medical Forms by Yourself): Not at all *(Reference)*	663 (14.62%)	--
A little bit	352 (7.76%)	1.16 (0.82, 1.65)
Somewhat	524 (11.55%)	**1.73 (1.27, 2.36)**
Quite a bit	853 (18.81%)	**2.10 (1.58, 2.78)**
Extremely	2144 (47.27%)	**2.04 (1.58, 2.64)**
Back Pain: None *(Reference)*	1754 (38.67%)	--
Very mild	1074 (23.68%)	0.85 (0.70, 1.02)
Moderate	1156 (25.49%)	0.85 (0.71, 1.03)
Fairly severe	391 (8.62%)	1.03 (0.78, 1.37)
Very severe or worst imaginable	161 (3.55%)	**1.70 (1.06, 2.71)**
Pain in Non-Operative Lower Extremity Joint:		

Variable	Frequency	OR (95% CI) (Weighted for Non-Response)
None *(Reference)*	1552 (34.22%)	--
Mild	1125 (24.80%)	0.85 (0.70, 1.03)
Moderate	1079 (23.79%)	0.93 (0.76, 1.14)
Severe	635 (14.00%)	**1.38 (1.07, 1.77)**
Extreme	145 (3.20%)	**1.97 (1.23, 3.18)**
BMI	*Mean*=30.46 (*SD* 6.03)	1.00 (0.99, 1.01)
Narcotic Use for >90 days	787 (17.35%)	0.97 (0.79, 1.18)
Baseline PROMIS Global Mental Health Score	*Mean*=49.70 (*SD* 8.05)	**0.98 (0.97, 0.99)**
Severe infection; other infectious diseases (CC 1, 3-7)	842 (18.56%)	0.94 (0.78, 1.13)
Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)	1217 (26.83%)	0.83 (0.59, 1.18)
Liver disease (CC 27-31)	1229 (27.09%)	1.22 (0.86, 1.73)
Rheumatoid arthritis and inflammatory connective tissue disease (CC 40)	457 (10.07%)	**0.78 (0.62, 0.99**)
Depression (CC 61)	698 (15.39%)	0.91 (0.73, 1.14)
Other psychiatric disorders (CC 63)	714 (15.74%)	0.96 (0.78, 1.20)
Coronary atherosclerosis or angina (CC 88-89)	1138 (25.09%)	**0.75 (0.64, 0.89)**
Vascular or circulatory disease (CC 106-109)	862 (19.00%)	0.92 (0.77, 1.12)
Renal failure (CC 135-140)	621 (13.69%)	1.08 (0.87, 1.34)

Table 6. Model Performance: Combined THA/TKA, THA-specific and TKA-specific Risk Models for SCB Improvement (Development Dataset)

Model Performance Statistic	Combined THA/TKA model	THA-specific model	TKA-specific model
C-statistic	0.68	0.68	0.68
Predictive Ability	26% – 81%	32% - 83%	25% - 81%

Figure 2. RSIRs Calculated using the THA-Specific (Hip Only) Model vs. RISRs Calculated using the Combined THA/TKA Model (Development Dataset)

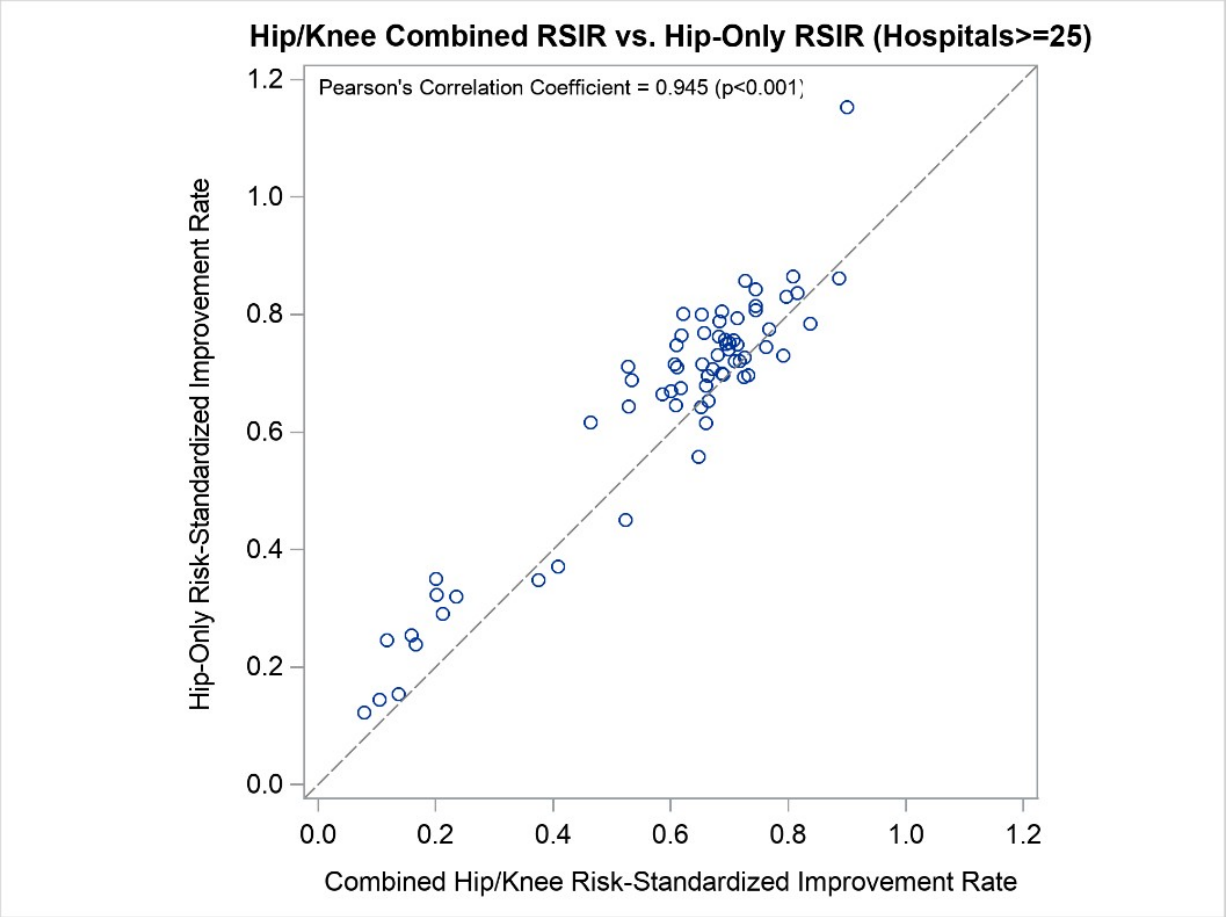
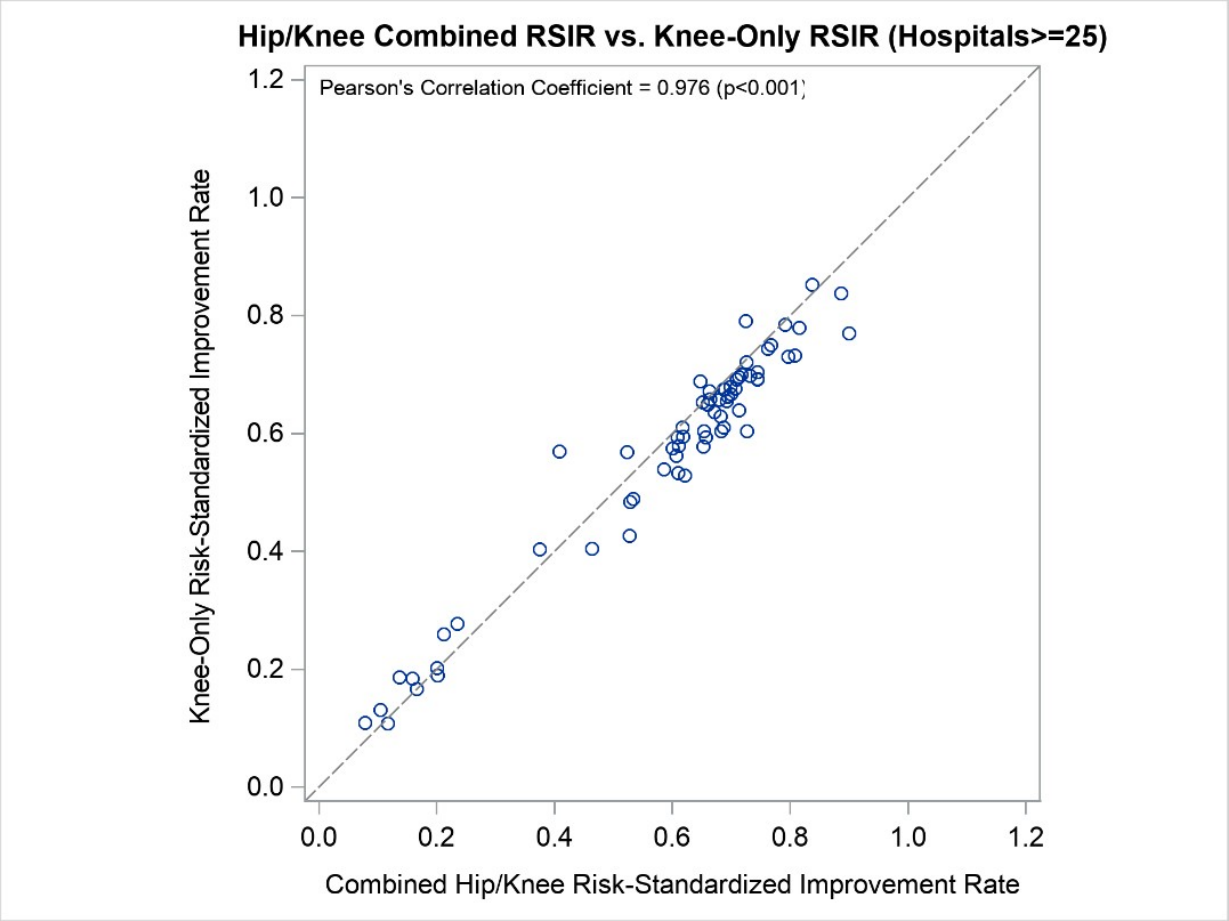


Figure 3. RSIRs Calculated using the TKA-Specific (Knee Only) Model vs. RISRs Calculated using the Combined THA/TKA Model (Development Dataset)



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.**

To explore the impact of social risk factors (in addition to health literacy, already included in the risk model), we examined the associations of dual eligibility and AHRQ SES Index lowest quartile (low SES) among patients undergoing primary elective THAs/TKAs with the measure outcome (SCB in PRO scores following surgery), using the Development Dataset. Due to known associations between race and poorer outcomes, we also assessed the association between non-White race and the outcome. Bivariate and multivariate analyses showed no statistically significant association between AHRQ SES Index lowest quartile and SCB improvement, nor non-White race and SCB improvement; dual eligibility was borderline significant (p=0.058) at the bivariate level (see Table 7 below), and statistically significant when entered into the risk model, indicating that patients with dual eligibility had higher odds of achieving SCB improvement (see Table 8 below). Table 9 provides the mean and range of hospital-specific RSIRs with no social risk factors included in the risk model, and with dual eligibility, and AHRQ SES Index lowest quartile individually included in the risk model. Correlation coefficients between RSIRs calculated without social risk factors with RSIRs calculated individually for each of the social risk factors indicates near perfect or perfect correlation in our data. This was also true when comparing RSIRs calculated without social risk factors with RSIRs calculated including non-White race. The lack of association and impact of these factors may be due to lower case selection in these groups for these elective primary procedures.

Based on the results of the social risk factor testing, we did not include additional social risk factors beyond health literacy. As noted above, we do include health literacy in the final risk model, based

upon strong patient and technical expert input. In our dataset, only dual eligibility was statistically significantly associated with the outcome, and while patients with dual eligibility had higher odds of achieving SCB improvement, inclusion of dual eligibility in the risk model did not appear to impact RSIRs. Additional analysis of hospital proportion of dual eligible patients by hospital RSIRs is provided in Figure 4. The results indicate that hospitals with the lowest proportion of dual eligible patients and those hospitals with the highest proportion of dual eligible patients have similar RSIR distributions. These data do not provide evidence of significant differences in RSIRs due to the proportion of a hospital's patients with dual eligibility.

Given this, we did not include additional social risk factors in the final risk model, beyond health literacy. However, we did find that social risk factors were significantly associated with response and therefore, we included social risk in our non-response adjustment of the measure (see Section 2b6 below). As this measure assesses patients undergoing an elective procedure where known disparities exist, we will continue to assess the impact of social risk for this measure over time.

Table 7. Bivariate Associations of Social Risk Factors and Race with SCB Improvement: Development Dataset (Patient N = 6,734, Hospital N = 230)

Variable	Frequency (%) of Total	Frequency (%) of Patients Achieving SCB Improvement	Frequency (%) of Patients Not Achieving SCB Improvement	P-value
Dual Eligibility	206 (3.06%)	146 (3.35%)	60 (2.52%)	0.0580
AHRQ SES Index: Lowest Quartile	688 (10.22%)	446 (10.24%)	242 (10.17%)	0.9222
Race: Non-White	548 (8.14%)	351 (8.06%)	197 (8.28%)	0.7569

Table 8. Adjusted Odds Ratios (ORs) for Social Risk Factors and Race Individually Evaluated in the Risk Model for SCB Improvement: Development Dataset (Patient N = 6,734, Hospital N = 230)

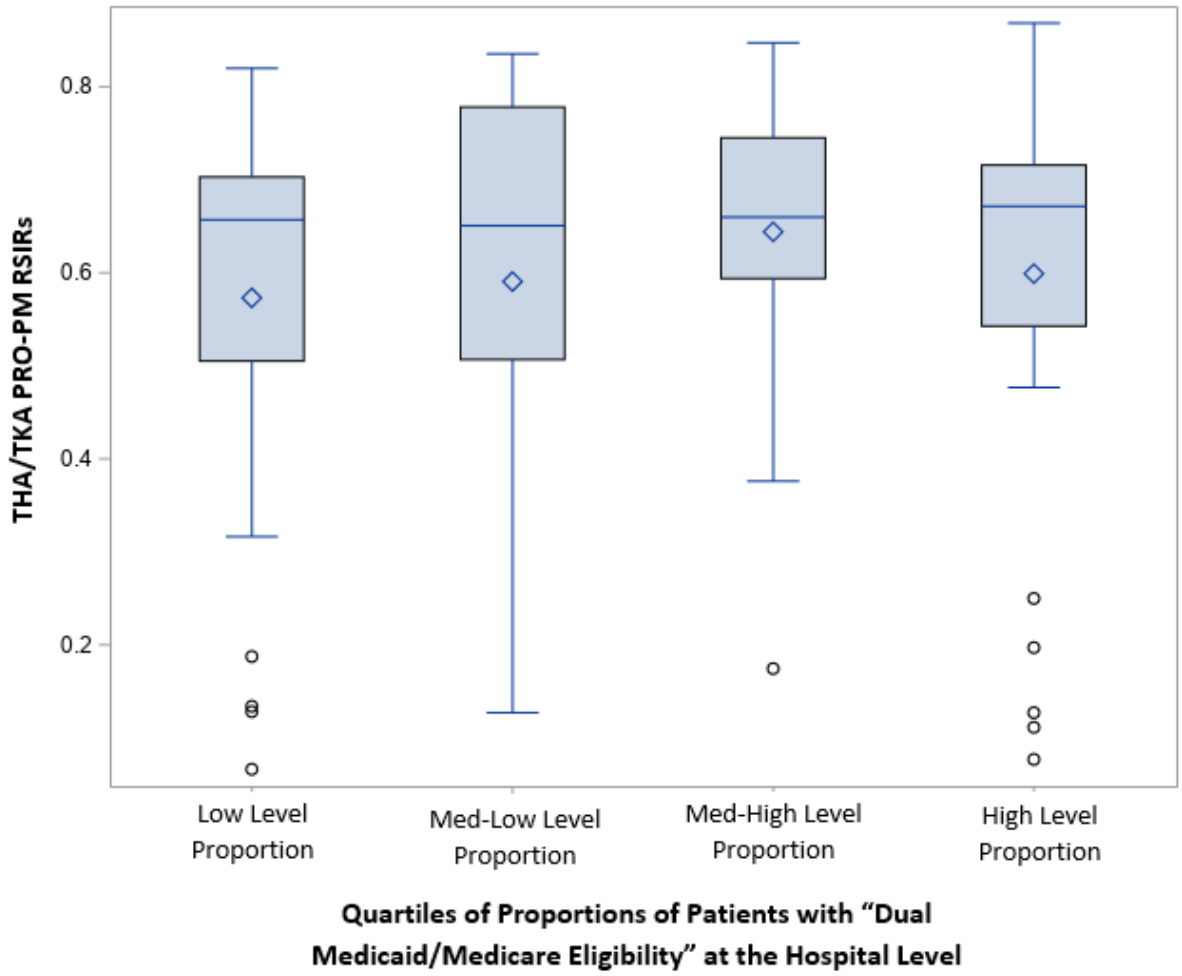
Variable	Frequency (%)	Estimate (SE)	OR (95% CI)	C Statistic for Model Including Social Risk Factor
Dual Eligibility	206 (3.06%)	0.40 (0.17)	1.49 (1.07, 2.08)	0.68*
AHRQ SES Index: Lowest Quartile	688 (10.22%)	0.04 (0.09)	1.04 (0.87, 1.25)	0.68*
Race: Non-White	548 (8.14%)	-0.08 (0.10)	0.93 (0.76, 1.13)	0.68*

** C-statistic for the risk model for SCB improvement in the Development Dataset without any of the three social risk factors = 0.68*

Table 9. Mean and Distribution of RSIRs Calculated without and with Social Risk Factors and Race in the Risk Model (Development Dataset: Hospitals with ≥25 THA/TKA Patients with PRO Data)

Summary Statistics	No Risk Factors Included	Dual Eligibility	AHRQ SES Index: Lowest Quartile	Race: Non-White
N (Hospitals)	94	94	94	94
Mean (SD)	60.39% (19.85)	60.40% (19.85)	60.30% (19.86)	60.36% (19.87)
Percentile				
100% Max	86.25%	86.21%	86.23%	86.03%
99%	86.25%	86.21%	86.23%	86.03%
95%	81.94%	81.96%	82.03%	81.71%
90%	79.95%	79.95%	79.95%	80.10%
75% (Q3)	72.37%	72.38%	72.33%	72.45%
50% (Median)	66.57%	66.53%	66.57%	66.60%
25% (Q1)	53.22%	53.23%	53.22%	53.26%
10%	20.07%	20.08%	20.06%	20.04%
5%	14.47%	14.49%	14.50%	14.43%
1%	8.47%	8.48%	8.46%	8.42%
0% Min	8.47%	8.48%	8.46%	8.42%
Pearson Correlation Coefficient (With "No Social Risk Factors")		0.9999	>0.9999	0.9997

Figure 4. THA/TKA PRO-PM RSIRs by Quartiles of Hospitals Grouped by Proportion of Dual Eligible Patients



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*)

To assess Model Performance, we computed discrimination and calibration statistics for assessing model performance (Harrell and Shih, 2001) for the clinically derived models, including:

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

(2) Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; good discrimination indicated by a wide range between the lowest decile and highest decile); and

(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). A value of close to zero for the intercept and close to 1 for coefficient of risk score indicates good calibration of the model.

Reference:

Harrell FE, Shih Y-CT. Using full probability models to compute probabilities of actual interest to decision makers. *International journal of technology assessment in health care.* 2001;17(1):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*)**:**

Model performance statistics for the risk model for meeting or exceeding the SCB improvement threshold are provided in Table 10 (below).

For the Development Dataset:

- **C-statistic for the risk model is 0.68**
- **Predictive ability from the lowest to highest decile is 26% - 82%**

For the Validation Dataset:

- **C-statistic for the risk model is 0.69**
- **Predictive ability from the lowest to highest decile is 26% - 81%**

Table 10. Model Performance of Risk-Adjusted Model of SCB Improvement following THA/TKA

Model Performance Statistic	Development Dataset	Validation Dataset
C-statistic	0.68	0.69
Calibration (γ0, γ1)	0.00, 1.00	-0.08, 1.02
Predictive Ability	26% – 82%	26% – 81%

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

The *calibration indices* (γ0, γ1) used to assess the risk model for meeting or exceeding SCB improvement are provided for the Validation Dataset in Table 10 (above): (-0.08, 1.02).

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

Figure 5 plots risk deciles for the Development Dataset; Figure 6 plots risk deciles for the Validation Dataset.

Figure 5. Calibration Deciles (Development Dataset)

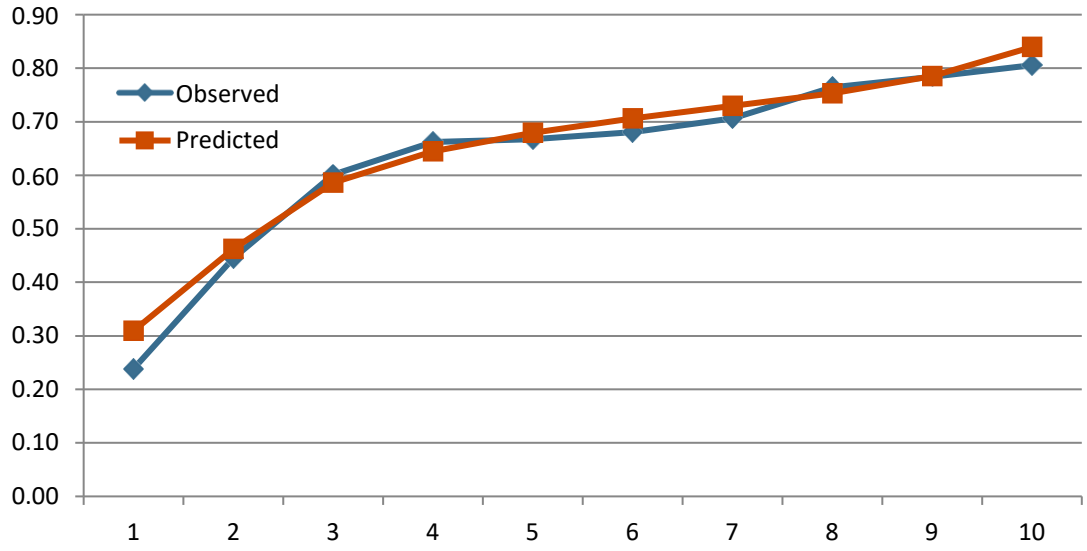
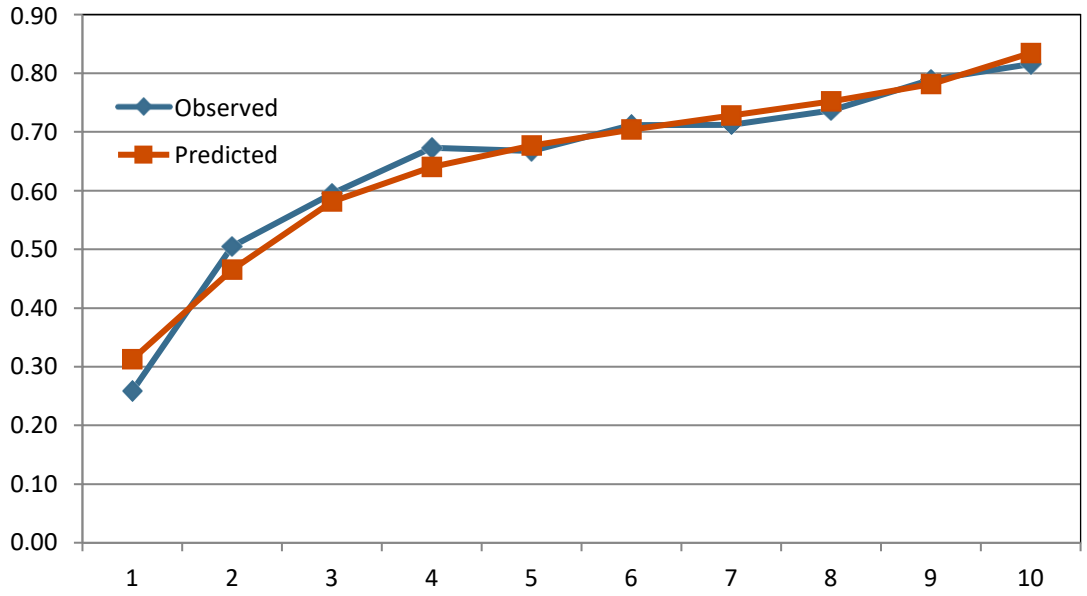


Figure 6: Calibration Deciles (Validation Dataset)



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*)

The following results demonstrate that the risk-adjustment model adequately controls for differences in patient characteristics: **<u>Discrimination statistics</u>** **The calculated c-statistic was 0.68 using the Development Dataset and 0.69 using the Validation Dataset and indicates adequate model discrimination across the cohort models. With both the Development and Validation Datasets, 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.** **<u>Calibration statistics (γ0, γ1)</u>** **The calibration values which are consistently close to 0 at one end and close to 1 at the other end indicates good calibration of the model. If the γ0 in the model performance using Validation data is substantially far from zero and the γ1 is substantially far from 1, there is potential evidence of over-fitting. The calibration values of close to zero at one end and close to 1 on the other end indicates good calibration of the model between the Development and Validation Datasets.** **<u>Risk Decile Plots</u>** **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.** **<u>Overall Interpretation</u>** **Interpreted together, our diagnostic results demonstrate that the risk-adjustment model adequately controls for differences in patient characteristics (case mix) and bias due to non-response.**

2b3.11. Optional Additional Testing for Risk Adjustment (*<u>not required</u>, 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)

Meaningful differences in performance measure scores are assessed by calculating the distribution of hospital-level RSIRs. Variation in hospital-level RSIRs indicate a clinically meaningful quality gap in the delivery of care to patients undergoing, as some hospitals can achieve substantially higher rates than the average performer, while other hospitals performing much worse than an average performer.

In addition, statistically significant differences were assessed using a median odds ratio (MOR) (Merlo et al, 2006). The median odds ratio represents the median increase in odds of the patient outcome (a SCB improvement in PROM score from preoperative to postoperative assessment) if a procedure on a single patient was performed by a higher performing hospital compared to a lower performing hospital. It is calculated by taking all possible combinations of hospitals (n=238 hospitals in the total dataset), always comparing the higher performing hospitals to the lower performing hospitals. The MOR is interpreted as a traditional odds ratio would be.

Reference:

Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, et al. (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, 60:290-297.

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*)

Table 11 provides the mean and distribution of hospitals' RSIRs. Risk-standardized improvement rates ranged from 6.65% to 86.84% (median: 66.49%).

Median Odds Ratio (MOR) = 3.44 with upper and lower 95% confidence bands of 3.385 and 3.485.

Table 11. Mean and Distribution of RSIRs for Risk Model of SCB Improvement following Elect6ive Primary THA/TKA (Hospitals with ≥25 THA/TKA Patients with PRO Data)

Summary Statistics	RSIRs (Combined Dataset)
N (Hospitals)	123
Mean (SD)	60.16% (19.58)
Percentile	
100% Max	86.84%

Summary Statistics	RSIRs (Combined Dataset)
99%	84.73%
95%	81.92%
90%	78.85%
75% (Q3)	72.51%
50% (Median)	66.49%
25% (Q1)	54.36%
10%	20.94%
5%	13.42%
1%	7.70%
0% Min	6.65%

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 variation in RSIRs (Table 9) suggests that there are meaningful differences in performance measure scores across hospitals. The interquartile range represents a difference of 18 percentage points, and the difference between the 10th and 90th percentiles (20.94% and 78.85%, respectively) is just shy of 58 percentage points. This variation indicates an important quality gap among hospitals.

The median odds ratio (MOR) suggests significant and substantial increases in the likelihood of SCB improvement by higher performing hospitals compared to lower performing hospitals. At the hospital level, the MOR value indicates that a patient is 3.44 times more likely to achieve SCB improvement if their elective primary THA/TKA procedure was performed by a higher performing hospital than by a lower performing hospital.

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 nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

Due to the voluntary nature of PRO survey data, we understand that accounting for potential non-response bias is important for this measure. With a thorough literature search, we identified several approaches for missingness (covariates adjustment in regression, submission score adjustment in regression, and stabilized inverse propensity score weighted regression). Following consultation with a statistical expert (Sharon-Lise Normand, PhD), we decided on addressing potential response bias using stabilized inverse probability weighting, as it would not modify the clinical risk model, and would not assume the form of a relationship between submission score and outcome (as suggested by Garrido 2016; Thoemmes and Ong 2016).

For this approach, we performed the following steps:

1) **All eligible THA/TKA procedures performed during the measurement period at the 238 hospitals submitting complete PRO and risk variable data for at least one of these procedures were identified via CMS claims data ("complete PRO submission," N=39,356 procedures).**
2) **These eligible THA/TKA procedures were categorized into one of three PRO response groups:**
   a) **Procedures for which complete PRO and risk variable preoperative data and complete PRO postoperative data were submitted ("complete PRO submission," N=11,270).**
   b) **Procedures for which incomplete PRO and risk variable data were submitted (including submissions with missing data elements and submissions of only preoperative PRO data or only postoperative PRO data ("incomplete PRO submission," N=10,133).**
   c) **Procedures for which no PRO data were submitted ("no response," N=17,953).**
3) **We compared patient characteristics and clinical comorbidities across the three PRO response groups and determined there were statistical differences in case-mix.**

4) **We conducted a literature review and identified the following variables associated with unit non-response to PROM survey data that were also available in our data: age, sex, race, low socioeconomic status, and post-operative complication following hip or knee procedures (Hutchings et al, 2012; de Rooij et al, 2018); Patel et al, 2015; Schamber et al, 2013).**
5) **Additional variables associated with PRO submission in our data were identified through multinomial logistic stepwise regression.**
6) **Propensity scores were calculated using a multinomial logistic regression where the outcome was 1) complete PRO submission, 2) incomplete PRO submission, and 3) no response.**
7) **Stabilized Inverse Probability Weights (IPW) were calculated for each of the three groups. For the complete responders, the stabilized weights were calculated using the following formula:** $\frac{PP(ZZ=11)}{PP(ZZ=11|xx)}$ **where** $(ZZ = 11)$ **represents the complete responders. Stabilized weights produce estimates with smaller variance and less extreme values compared to using the standard non-stabilized weights calculated in the following way:** $\frac{11}{PP(ZZ=11|xx)}$**. Table 12 provides the distribution of the stabilized weights with mean 1.00 and standard deviation of 0.26.**
8) **The stabilized IPW were incorporated into the hierarchical risk-adjustment model for SCB improvement following elective primary THA/TKA and used in calculation of the risk-adjusted and bias-adjusted RSIRs.**

Incorporating the stabilized weights in the calculation of the RSIRs helps to reduce bias due to non-response by giving higher weight to patients who were less likely to respond and deflating the weight of patients who were more likely to respond based on patient characteristics. Weighting the responders based on their likelihood of response, given their patient characteristics, helps reduce non-response bias in our RSIR measure.

Among the 238 hospitals submitting at least one complete PRO submission for an eligible THA/TKA procedure during the measurement period, 389 (0.89%) patients died before having the opportunity to complete postoperative PRO data. Given the small number of deaths, we excluded those who died within 9 months of the procedure from the propensity score model.

References:

Hutchings A, Neuburger J, Frie KG, Black N, van der Meulen J. (2012). Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England. Health and Quality of Life Outcomes, 10, 34 doi:10.1186/1477-7525-10-34;

de Rooij BH, Ezendam NPM, Mols F, Vissers PAJ, Thong MSY, Blooswijk CCP, Oerlemans S, Husson O, Horevoorts NJE, van de Poll-Franse LV. (2018). Cancer survivors not participating in observational patient-reported outcome studies have a lower survival compared to participants: the population-based PROFILES registry. Quality of Life Research, 27:3313-3324.

Garrido, M. M. (2016). Covariate Adjustment and Propensity Score. Jama, 315(14), 1521. doi: 10.1001/jama.2015.19081

Patel J, Lee JH, Zhongmin L, SooHoo NF, Bozic K, Huddleston JI. (2015). Predictors of low patient-reported outcomes response rates in the California Joint Replacement Registry. The Journal of Arthroplasty, 30:2071-2075.

Thoemmes, F., & Ong, A. D. (2015). A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. Emerging Adulthood, 4(1), 40–59.

Schamber EM, Takemoto SK, Chenok KE, Bozic KJ. (2013). Barriers to completion of patient reported outcome measures. The Journal of Arthroplasty, 28:1449-1453.

Table 12: Distribution of Stabilized Weights Applied to Patients with Complete PRO Submission (Responders)

Summary Statistics	Stabilized Weights
Mean (SD)	1.00 (0.26)
Percentile	
100% Max	4.74
99%	1.77
95%	1.29
90%	1.09
75% (Q3)	1.01
50% (Median)	0.95
25% (Q1)	0.91
10%	0.88
5%	0.85
1%	0.82
0% Min	0.73

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*)

Patients included in measure development and testing of this measure had complete preoperative PRO and risk variable data matched to complete postoperative PRO data. Patients with PRO submissions that were incomplete: missing data values, data values out-of-range, or missing preoperative or postoperative PRO data were not included in the Development and Validation Datasets.

The true "response" rate for our study is difficult to calculate because it is unknown to whether 100% of eligible patients at the hospitals in our dataset were asked to provide PRO data. However, we do have the true denominator of eligible cases, based upon claims data. In the absence of a true "response" rate, we have calculated an estimated response rate as the percentage of all elective

primary THA/TKA procedures performed during the measurement period at the hospitals in the dataset (excluding patients with staged procedures during the measurement period) for which complete and matched preoperative and postoperative PRO and risk variable data were submitted. With this operational definition, the mean response rate across hospitals was 30.62% (*SD* 22.79%). Among hospitals with ≥25 elective primary THA/TKA patients with PRO data during the one-year measurement period, the mean response rate among hospitals was 43.15% (See Table 13, below). The CJR model required either a minimum percentage or an absolute minimum number of PRO cases be submitted to qualify for the quality point incentive; the thresholds in CJR performance years one and two were 50% of or 50 eligible cases and 60% of or 75 eligible cases, respectively.

To address potential response bias using stabilized inverse probability weighting, created with a multinomial logistic regression to calculate stabilized inverse probability weights. We checked for model fit of the propensity score model by Hosmer Lemeshow test for goodness of fit and did not find evidence of lack of fit (Hosmer-Lemeshow statistic was 12.06, p-value = 0.15).

Results of the stabilized inverse probability weighting to address potential non-response bias are reflected in the comparison of mean and distribution of hospital RSIRs for risk-adjusted model of SCB improvement with and without stabilized inverse probability weighting (Table 14, below).

Table 13. Mean and Distribution of Hospital Response Rates (for Complete PRO and Risk Variable Data, Combined Dataset)

Summary Statistics	Response Rates (All Hospitals)	Response Rates (Hospitals with ≥25 THA/TKA Patients with PRO Data)
N (Hospitals)	238	123
Mean (SD)	30.62% (22.79)	43.17 (20.52)
Percentile		
100% Max	100.00%	90.50%
99%	84.78%	89.66%
95%	74.29%	79.64%
90%	61.45%	69.66%
75% (Q3)	46.23%	60.58%
50% (Median)	27.88%	40.85%
25% (Q1)	9.68%	28.34%
10%	3.70%	17.74%
5%	2.06%	11.49%
1%	0.72%	5.65%
0% Min	0.24%	5.00%

Table 14. Mean and Distribution of Hospital RSIRs for Risk-Adjusted Model of SCB Improvement With and Without Stabilized Inverse Probability Weighting for Potential Non-Response Bias (Combined Dataset, Hospitals with ≥25 THA/TKA Patients with PRO Data)

Summary Statistics	Risk-Standardized Improvement Rates (No Weighting)	Risk-Standardized Improvement Rates (Weighted for Non-Response)
N (Hospitals)	123	123
Mean (SD)	60.21% (19.57)	60.16% (19.58)
Percentile		
100% Max	86.66%	86.84%
99%	85.34%	84.73%
95%	81.69%	81.92%
90%	78.98%	78.85%
75% (Q3)	72.77%	72.51%
50% (Median)	66.18%	66.49%
25% (Q1)	54.63%	54.36%
10%	21.70%	20.94%
5%	13.19%	13.42%
1%	7.79%	7.70%
0% Min	6.89%	6.65%

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 nonresponders) and how the specified handling of missing data minimizes bias**?** (i*.e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)

We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation is 0.00194 (p-value=0.84). This indicates that there is not an association between the residuals and the probability of response based on our model.

We examined the correlation between the residuals of the stabilized inverse probability weighted hierarchical model and the submission probability finding it to be 0.00492 (p-value=0.60) suggesting that there is not an association between the residuals weighting for non-response and probability of response.

The correlation between RSIR unadjusted and inverse probability weighted RSIR is very high suggesting that the results are not sensitive to our weighting adjustment. However, due to the high proportion of non-responders, we considered it important to account for the differences in characteristics of responders and non-responders found in the literature and empirically in our data.

We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation is 0.00194 (p-value=0.84). This indicates that there is not an association between the residuals and the probability of response based on our model.

The comparison of hospital RSIRs for risk-adjusted model of SCB improvement with stabilized inverse probability weighting and without stabilized inverse probability weighting (Table 13, above) reveals only a small impact on the measure results of adjusting for potential non-response. However, we expect that non-response bias will be a factor for the THA/TKA PRO-PM measure, due to associations with non-response including socioeconomic status and health status. We therefore retained response bias adjustment for the measure results.

3. Feasibility

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

3a. Byproduct of Care Processes

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

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

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

If other:

3b. Electronic Sources

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

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

No data elements are in defined fields in electronic sources

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

Currently, this measure allows hospitals to collect data using a range of methods, including paper and electronic formats. While we strongly support the use of electronic data capture, not all clinicians collect patient-reported outcomes on their patients eligible for and undergoing elective primary THA/TKA procedures and many fewer collect these data in electronic form. In fact, the vast majority of hospitals participating in the Center for Medicare and Medicaid Innovation (CMMI) Comprehensive Care for Joint Replacement (CJR) model submitting PRO data do not use electronic data capture. The rapid and continual advances being made in mobile applications and other modes of electronic PRO data capture support likely feasibility of moving to an electronic format for this measure in the near future in ways that were not available at the time of measure development. Further the specifications are harmonized with eCQM process measures that incentivize

collection of the PRO data needed to calculate the measure outcome, making future e-specification less burdensome.

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.

Although PROMs are not universally collected prior to and following THA and TKA procedures, incentivized PRO data collection within CMS's Comprehensive Care for Joint Replacement (CJR) model presents proof of concept for feasible, low burden collection of PROs for hospital-level quality measurement. Challenges to PRO collection can be mitigated by strong leadership support, flexibility in rearranging clinical workflows to accommodate PRO data collection, ability to access PRO data in real-time for clinical decision making, and universal staff buy-in on the value of PROs in improving care and quality.

Some amount of missing data and non-response may be expected given the voluntary nature of PRO data, even with the above approaches. Therefore, the statistical methods use stabilized inverse probability weighting (IPW) to address potential non-response bias.

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)

Payment Program Not in use	

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

N/A; this PRO-PM is being submitted for initial endorsement and is not currently used in any accountability program.

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?*)

This PRO-PM is being submitted for initial endorsement and is not currently used in any accountability program.

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.*)

This PRO-PM will be implemented in to-be-determined federal accountability programs through rulemaking in the future.

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.

This PRO-PM has not been implemented yet and thus measure results have not been shared with the measured entities (hospitals). However, feedback was obtained from a TEP (23 total members, five of which were patients), a Technical Advisory Group (eight members), and a Patient Working Group (six total members). These individuals were selected through a publicly posted call for TEP members on the CMS website or through partnerships with the National Partnership for Women and Families and Rainmakers. Feedback was obtained via teleconference calls and online surveys. Patients engaging in this work were provided with preparation calls that reviewed the meeting materials ahead of the meeting date and debrief calls that allowed them to share any thoughts after the scheduled meeting. All meeting materials were sent in advance to allow individuals time to review the performance results and data. A summary of the feedback is provided in Section 1a.3 (value and meaningfulness) of the NQF Evidence Form.

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.

The Technical Expert Panel (TEP) has been engaged in measure development since the conceptual stage. They have provided input on cohort, outcome, and risk adjustment decisions. The Technical Advisory Group was consulted on determining an outcome and selection of Patient Reported Outcome Measures (PROMs). The Patient Working Group provided input on measure outcome, risk adjustment, and testing results. Statistical analyses were shared with the TEP and Patient Working Group.

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.

Feedback was obtained via seven teleconference meetings with the TEP, three teleconference meetings with the Patient Working Group, and one online survey administer to the Technical Advisory Group.

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

Measure results have not been shared with hospitals, but the TEP, which had multiple clinicians indicated strong support for a patient-reported outcomes performance measure following elective THA and TKA.

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

The Patient Working Group members indicated strong support for a patient-reported outcomes performance measure following elective THA and TKA. Patients expected a significant amount of improvement in pain levels and functional status. Patients noted that the procedure impacted their physical health and their quality of life, and find the measure to be valuable.

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.

TEP, Technical Advisory Group, and Patient Working Group feedback has been considered in the development of this measure through the selection of a cohort, measure outcome, data collection instruments, and risk adjustment models. Patients provided input on the amount of change they would like to see, which helped define the thresholds for the measure outcome.

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.

This is a new PRO-PM, not currently used in a quality improvement program, and there are no performance results to assess. A primary goal of the PRO-PM following implementation in a federal accountability program is to provide hospitals with performance information necessary to implement focused quality improvement efforts.

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.

N/A; this is a new PRO-PM not yet implemented. No unexpected findings were noted during PRO-PM development or testing.

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

N/A; this is a new PRO-PM not yet implemented. No unexpected benefits were noted during PRO-PM development or testing.

5. Comparison to Related or Competing Measures

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

5. Relation to Other NQF-endorsed Measures

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

Yes

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

0422 : Functional status change for patients with Knee impairments

0423 : Functional status change for patients with Hip impairments

0424 : Functional status change for patients with Foot and Ankle impairments

0425 : Functional Status Change for Patients with Low Back Impairments

0426 : Functional status change for patients with Shoulder impairments

0427 : Functional status change for patients with elbow, wrist and hand impairments

0428 : Functional status change for patients with General orthopaedic impairments

1550 : Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA)

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

2643 : Average change in functional status following lumbar spine fusion surgery

2958 : Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery

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

NQF # 2653: Average change in functional status following total knee replacement surgery

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.

To the extent feasible, we have harmonized with existing, related measures. However, we have prioritized the goal of the measure to assess substantial clinical benefit (SCB) improvement in patient-reported outcomes for elective primary THA/TKA patients with minimal patient and provider burden over harmonization if discrepancies occur.

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.)

NQF # 2653: Average change in functional status following total knee replacement surgery.
This PRO-PM measure differs from NQF #2653 in attribution, cohort, outcome, and risk adjustment.
Attribution: This PRO-PM is a hospital-level quality measure, whereas NQF #2653 is a clinician-level measure.
Cohort: This PRO-PM includes both THA and TKA procedures, as clinical experts agree that hospital-level processes are shared across these procedures, and includes only primary, not revision, procedures, based upon clinical input that revision procedures are more complicated to perform and patient-reported outcomes may be influenced by the initial surgery. The target population is Medicare FFS beneficiaries 65 years of age and older. NQF #2653 includes only TKA procedures, includes knee replacement revisions as well as primary procedures, and includes all adults 18 years of age and older.
Outcome: This PRO-PM collects PROs with the HOOS, JR for THA patients and the KOOS, JR for TKA patients. Timing of PRO data collection is 90 – 0 days prior to and 270 – 365 days following surgery. The numerator measures SCB improvement for each patient from preoperative to postoperative assessment with a binary outcome (Yes/No), and the measure produces a risk-standardized improvement rate that elucidates for hospitals the risk-adjusted proportion of patients with improvement and those without improvement. In contrast, NQF #2653 collects PRO data with the Oxford Knee Score three months prior to and 9 – 15 months following surgery, and measures average change in knee function score. The outcome definition of SCB, with a defined threshold for change in PROM score, allows patients with poorer baseline PRO scores more room to improve and thus a greater opportunity to achieve SCB. This was identified by our TEP members as a specific benefit of measuring SCB versus average change; measuring SCB incentivizes providers to offer and perform THA/TKA procedures on even those with poor PRO scores. Further stated TEP and Patient Working Group concerns with measuring an average change score included the fact that hospitals with all average outcomes would look similar to hospitals whose patients either did very well or very poorly (bimodal distributed outcomes), thus providing potentially misleading information to consumers and patients.
Risk Adjustment: This risk model for this PRO-PM includes important risk variables supported by technical expert panel (TEP) and other expert clinical consultants including health literacy, other musculoskeletal pain and chronic narcotic use which are not included in NQF #2653; these risk variables were identified and tested based upon input from orthopedic professional societies, including AAHKS and AAOS, through public comment (Centers for Medicare & Medicaid Services , CJR Final Rule 2015, Section III.D.3.A).
This PRO-PM is superior to NQF #2653: 1) it more appropriately provides a signal of hospital quality which reflects outcomes for both THA and TKA recipients since within hospitals, care for patients undergoing THA/TKA procedures is provided by the same providers and hospital staff; 2) it assesses SCB improvement with a binary outcome that elucidates for hospitals and patients the risk-adjusted proportion of patients with and without improvement (a clear, understandable metric that patients support); 3) it uses a more robust and stakeholder-driven risk model, anticipated to produce a measure with greater face validity with stakeholders; and 4) it is harmonized with related measures including NQF #1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and total knee arthroplasty (TKA) and 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 (MUC19-28).
References:
Comprehensive Care for Joint Replacement (CJR) Payment Model for Acute Care Hospitals Furnishing Lower Extremity Joint Replacement Services Final Rule, 80 C.F.R. 73273 (Nov 24, 2015).

Appendix

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

Attachment:

Contact Information

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

Co.2 Point of Contact: Vinitha, Meyyur, Vinitha.meyyur@cms.hhs.gov, 410-786-8819-

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: Lisa, Suter, lisa.suter@yale.edu, 203-764-5700-

Additional Information

Ad.1 Workgroup/Expert Panel involved in measure development

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

Yale New Haven Health Services Corporation/Center for Outcomes Research (YNHHSC/CORE) Measure Team Members

1. Lisa Suter, MD –Associate Director, Quality Measurement Program. Provided experience relevant to clinical content and performance measurement.

2. Kathleen Balestracci, PhD – Lead. Provided experience relevant to performance measurement.

3. Zhenqiu Lin, PhD- Analytic Director. Provided experience relevant to performance measurement.

4. Sarah Zimmerman, MS – Lead Analyst. Provided experience relevant to performance measurement.

5. Yongfei Wang, MS – Supporting Analyst. Provided experience relevant to performance measurement.

6. Sheng Zhou, MD, ScM - Supporting Analyst. Provided experience relevant to performance measurement.

7. Kyaw Sint, PhD - Supporting Analyst. Provided experience relevant to performance measurement

8. Elina Kurkurina, MPH – Research Project Coordinator. Provided experience relevant to performance measurement.

9. Lynette Lines, MST, PMP – Research Project Manager. Provided experience relevant to performance measurement.

10. Julia McMahon, BS – Research Assistant II. Provided experience relevant to performance measurement.

11. Susannah Bernheim, MD, MHS- Director of CMS Hospital Measures, Clinical Investigator. Provided experience relevant to clinical content and performance measurement.

12. Harlan Krumholz, MD, SM- Director of CORE. Provided experience relevant to clinical content and performance measurement.

Technical Expert Panel (TEP) Members

1. Peter G. Allen, MS- Regulatory Scientist/Biomedical Engineer, Food and Drug Administration (FDA). Provided experience relevant to performance measurement.

2. David C. Ayers, MD- Professor of Orthopedics, University of Massachusetts (UMass) Medical School. Provided experience relevant to clinical content and performance measurement.

3. Thomas C. Barber, MD- Vice President of Perioperative Services at UCSF and Professor of Orthopedic Surgery. Provided experience relevant to clinical content and performance measurement.

4. Daniel J. Berry, MD- Chairman of Department of Orthopedic Surgery, Mayo Clinic. Provided experience relevant to clinical content and performance measurement.

5. Vinod Dasa, MD- Associate Professor, Department of Orthopaedic Surgery, Louisiana State University Health Sciences Center. Provided experience relevant to clinical content and performance measurement.

6. Cheryl Fahlman, PhD, MBA, BSP- President, CAF Consulting Solutions. Provided experience relevant to performance measurement.

7. Cynthia S. Jacelon, PhD, RN-BC, CRRN, FAAN- Association of Rehabilitation Nurses; Associate Professor, University of Massachusetts Amherst School of Nursing. Provided experience relevant to clinical content and performance measurement.

8. Courtland G. Lewis, MD- Director of Orthopedic Surgery, Hartford Hospital. Provided experience relevant to clinical content and performance measurement.

9. Patient – Recipient of elective THA or TKA procedure. Provided patient perspective.

10. Michael H. Perskin, MD- The American Geriatrics Society; Associate Chair of Clinical Affairs and Assistant Professor in the Department of Medicine, New York University School of Medicine. Provided experience relevant to clinical content and performance measurement.

11. Jonathan L. Schaffer, MD, MBA- Managing Director, eCleveland Clinic Information Technology Division of The Cleveland Clinic Foundation. Provided experience relevant to clinical content and performance measurement.

12. John H. Seiverd, MD, MBA- Physical Therapy Center Coordinator of Clinical Education, Orthopaedic and Neurologic PT Residency Program Director, James A. Haley Veterans' Hospital. Provided experience relevant to clinical content and performance measurement.

13. Lyle S. Sorensen, MD- Chief of Orthopedics and Sports Medicine, Virginia Mason Medical Center. Provided experience relevant to clinical content and performance measurement.

14. A. Christopher Strenta, PhD- Recipient of elective THA or TKA procedure; Associate Dean, Finance and Operations, Dartmouth College. Provided patient perspective.

15. Margaret A. VanAmringe, MHS- Vice President, Public Policy and Government Relations, The Joint Commission. Provided experience relevant to performance measurement.

16. Rachel DuPre Brodie – Director of Performance Information, Pacific Business Group on Health. Provided experience relevant to performance measurement.

17. Sandra Geisinger, RN, EdD – Recipient of elective THA or TKA procedure. Provided patient perspective.

18. Cherie Gress - Recipient of elective THA or TKA procedure. Provided patient perspective.

19. William Hamilton, MD – Clinical Instructor, Anderson Orthopaedic Clinic; Chair of the Quality Measures Committee, American Association of Hip and Knee Surgeons. Provided experience relevant to clinical content and performance measurement.

20. Arthur Malkani, MD - Chief of Adult Reconstruction and Clinical Professor, Department of Orthopedic Surgery, University of Louisville. Provided experience relevant to clinical content and performance measurement.

21. Nan Rothrock, PhD – Research Associate Professor. Provided experience relevant to performance measurement.

22. Adolph J. Yates, Jr, MD – Orthopaedic Surgeon/Associate Professor, University of Pittsburgh Medical Center. Provided experience relevant to clinical content and performance measurement.

23. Patient – Recipient of elective THA or TKA procedure. Provided patient perspective.

Technical Advisory Group

1. Inder Johnson, MBA, OTR/L – Director of Rehabilitation Services and Occupational Therapist, Adventist Health and Rideout Hospital. Provided experience relevant to clinical content.

2. Sheila Barnett, MD – Associate Professor of Anesthesiology, Beth Israel Deaconess Medical Center. Provided experience relevant to clinical content.

3. Brian Curtin, MD, MS – Orthopedic Surgeon, OrthoCarolina Hip and Knee Center. Provided experience relevant to clinical content.

4. Sarah R. Piva, PT, PhD – Associate Professor of Physical Therapy, University of Pittsburgh. Provided experience relevant to clinical content.

5. Jennifer Lennon, OTR/L – Director of Rehabilitation Services, UPMC Presbyterian, Montefiore, and Western Psychiatric Institute Clinic. Provided experience relevant to clinical content.

6. Brandy N. Wilkins, DTP – Program Coordinator, The Joint Commission. Provided experience relevant to performance measurement.

7. Paula Farrell, BSN, RN, CPHQ – Associate Project Director, The Joint Commission. Provided experience relevant to performance measurement.

8. Daniel Riddle, PT, PhD, FAPTA – Otto D. Payton Professor of Physical Therapy, Virginia Commonwealth University. Provided experience relevant to clinical content.

Work Group Member

1. Kevin Bozic, MD, MBA- William R. Murray Professor, Chair of the Department of Surgery and Perioperative Care, and Professor of Orthopaedic Surgery at the University of Texas at Austin Dell Medical School. Provided experience relevant to clinical content and performance measurement.

Technical Advisors

1. Kate Chenok, MBA - President of Chenok Associates. Prepared materials for stakeholder engagement.

Measure Developer/Steward Updates and Ongoing Maintenance

Ad.2 Year the measure was first released:

Ad.3 Month and Year of most recent revision:

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

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

Ad.6 Copyright statement:

Ad.7 Disclaimers:

Ad.8 Additional Information/Comments: