NATIONAL QUALITY FORUM
Driving measurable health improvements together

Measure Worksheet

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

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

Brief Measure Information

NQF #: 3639

Corresponding Measures:

Measure Title: Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM)

Measure Steward: Centers for Medicare & Medicaid Services

sp.02. Brief Description of Measure: This patient-reported outcome-based performance measure uses the same measure specifications as the NQF-endorsed (NQF # 3559) hospital-level risk-standardized improvement rate (RSIR) following elective primary THA/TKA with the following exception: this measure attributes the outcome to a clinician or clinician group. Specifically, this measure will estimate a clinician-level and/or a clinician group-level 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.01. Developer Rationale: The goal of this measure is to improve patient outcomes by providing information to patients and clinicians about clinician- and clinician group-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 surgical approach and technique, perioperative planning, shared decision making with the patient, 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 by the surgeon performing the THA or TKA procedure, but also by patient status on presentation, this measure is risk-adjusted to account for patient-level characteristics. THA/TKA procedures provide a particularly rich test bed for developing quality measures based upon patient-reported experiences and piloting performance measures based upon PROMs. 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 a PRO-based measure via TEP and Patient Working Group engagement. Likewise, the hospital-level THA/TKA PRO-PM upon which this measure is based had strong patient support.

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

---

sp.12. Numerator Statement: The numerator is the risk-standardized proportion of patients undergoing an elective primary THA or TKA who experience a 22 point or 20 point or more improvement, for hip replacement and knee replacement patients respectively between preoperative and postoperative assessments on joint-specific patient-reported outcome measures (PROMs). The patient-level improvement thresholds are an a priori, patient-defined substantial clinical benefit (SCB) threshold of improvement which is an anchor-based threshold developed using patient-report of satisfaction with change in Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR)/Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS JR) scores (Lyman and Lee, 2018). This measure uses the same SCB threshold developed for the hospital-level measure, which was reviewed and recommended for endorsement by the NQF Surgery Standing Committee in 2020. 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 the hospital-level THA/TKA PRO-PM development Patient Working Group, Technical Expert Panel (TEP), Technical Advisory Group, and Orthopedic Clinical Expert.

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.

sp.14. 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.

The cohort does not include patients with hip fractures, pelvic fractures, revision THAs/TKAs, and bone metastases. The rationale for each is outlined below:

- **Facture of the pelvis or lower limbs coded in the principal or secondary discharge diagnosis fields on the index admission claim** (Note: Periprosthetic fractures must be additionally coded as POA in order to disqualify a THA/TKA from cohort inclusion, unless exempt from POA reporting.) Rationale: Patients with fractures have higher mortality, complication, and readmission rates, and the procedures are typically not elective.
- **A concurrent partial hip or knee arthroplasty procedure** Rationale: Partial arthroplasty procedures are primarily done for hip and knee fractures and are typically performed on patients who are older, frailer, and have more comorbid conditions.
- **A concurrent revision, resurfacing, or implanted device/prosthesis removal procedure** Rationale: Revision procedures may be performed at a disproportionately small number of hospitals and are associated with higher mortality, complication, and readmission rates. Resurfacing procedures are a different type of procedure involving only the joint's articular surface and are typically performed on younger, healthier patients. Elective procedures performed on patients undergoing removal of implanted device/prostheses procedures may be more complicated.
- **Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field on the index admission claim** Rationale: Patients with these malignant neoplasms are at increased risk for complication, and the procedure may not be elective.

sp.16. Denominator Exclusions: The measure has three denominator exclusions, listed below.

1. Staged Procedures

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 from the measure cohort.

2. Patients who die within 270 days of the procedure

All patients who expired within 9 months (270 days) of the THA/TKA procedure are removed from the measure cohort.

3. Patients who leave against medical advice from the inpatient index admission

Finally, patients who leave their index admission against medical advice are removed from the measure cohort.

Please note that hospice patients should not be excluded from the measure cohort because any patient undergoing a major surgery such as THA/TKA most likely has short-term survival as the primary goal.

Please also note that patients without complete PROM data, such as those that refuse to complete the PROM, are excluded from the measure results given the measure requires complete PROM data to calculate the measure outcome. Patients with incomplete or no PROM data are included in the non-response bias adjustment to alleviate potential bias. Further, CMS is exploring reporting response rate or other information along with the measure results to provide the end user of the measure results with a better sense of the sample being assessed by the measure.

Below we answer additional questions from NQF staff regarding these exclusions:

Question 1, Staged Procedures:

Please explain how staged procedures are assessed when they overlap the end and beginning of measurement periods. Is there an acceptable range in days for a staged procedure? Are all staged procedures planned? Do all staged procedures need to occur in the inpatient/acute care setting? Is it possible to have 1 inpatient and 1 outpatient surgery on the same joint? Are these procedures staged? How does that impact the denominator?

CORE response: To clarify, a "staged procedure" is a bilateral THA or TKA (both right and left hips or both right and left knees). Bilateral THAs and TKAs can be performed at the same time (these are included in the measure cohort), or during separate hospitalizations (these are the excluded "staged procedures"). Therefore, all staged procedures are planned. Theoretically, a staged procedure could be performed in different settings (for example, right THA performed inpatient followed by a left THA performed in the outpatient setting), but our clinical advisors suggest this is currently rare, although it may increase in prevalence over time.

During measure development, we only assessed staged procedures as any subsequent elective, primary THA/TKA procedure in the inpatient setting that occurred during the measurement period. In the future, we will need to assess the feasibility of extending the assessment of staged procedures to before and/or after the measurement period. Of note, this exclusion represents a small number of the total patients undergoing THA and TKA procedures in our testing dataset.

Based on discussions with our orthopedic experts, including Dr. Kevin Bozic, many staged THA/TKA procedures occur within 6 months of each other; timing is solely dependent upon provider and patient discussion of the patient's unique situation and formal guidelines do not exist. We used the measurement period given the measure has approximately a year postoperative PRO data collection window and any procedure that occurs during the postoperative PRO data collection window may negatively impact the recovery of the first procedure and it may be challenging to distinguish the recovery for either procedure from the other when they occur within 12 months of each other. In our dataset, we found that 1,181 (91.4%) of staged procedures occurred within 1 year and 111 (8.6%) of staged procedures occurred within 2 years.

To qualify as a staged procedure in the measure, the procedure must meet the criteria of an elective primary procedure. Yes, the current cohort exclusion requires staged procedures to occur in the inpatient setting. In the future we will assess staged procedures that may occur in the outpatient setting (hospital outpatient departments and ambulatory surgical setting). In the example of 1 inpatient and 1 outpatient surgery on the same joint is unlikely a staged procedure, rather a revision or other non-elective procedure on the same joint. As noted above, this is not how we define "staged procedures". The measure cohort does not include revision procedures in measure cohort therefore subsequent procedures on the same joint that do not meet cohort criteria would not be included in the cohort.

Question 2: AMA exclusion

Are there any other forms of AMA that are appropriate for the measure, such as patients who "fire" their providers?

At this time, we only use the discharge disposition code to identify patients who leave AMA. In the example you provide of a patient "firing" their provider, please note that this information would not be systematically captured in claims data and therefore we would be unable to investigate these instances.

Measure Type: Outcome: PRO-PM

sp.28. Data Source: Claims, Instrument-based, Other (specify)

sp.07. Level of Analysis: Clinician: Group/Practice

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:

sp.03. IF PAIRED/GROUPED, what is the reason this measure must be reported with other measures to appropriately interpret results?

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 is a new, patient-reported outcome performance measure (PRO-PM) utilizing claims, instrument-based, and Medicare enrollment data at the individual clinician and group/practice level that aims to improve patient outcomes by providing information to patients and clinicians about clinician- and clinician group-level, risk-standardized patient-reported outcomes, such as pain and functional status, following elective primary THA/TKA.
- The logic model presented by the developer for this outcome measure links actions that can be taken by the accountable entity— such as surgical approach and technique, perioperative planning, shared decision making with the patient, communication among providers, prevention of and response to complications, patient safety, and coordinated transitions to the outpatient environment- with patient-reported outcomes (PROs) (i.e., improved recovery and rehabilitative status and decreased pain and improved mobility and quality of life) following total hip and/or total knee arthroplasty (THA/TKA).
- The developer highlighted those patients on both the developer's Technical Expert Panel and Patient Working Group indicated they found the measure to be meaningful. The developer noted evidence supports attributing patient-reported outcomes to the surgeons performing the procedure, including data supporting that low surgeon case volume is associated with longer operating times, lengthier hospitalizations, higher infection rates, and worse PROs.
- The developer noted supporting evidence that attributes patient-reported outcomes to the surgeons performing the procedure, including data supporting that low surgeon case volume is associated with longer operating times, lengthier hospitalizations, higher infection rates, and worse PROs.

Guidance from the Evidence Algorithm

Does the measure assess performance on a health outcome (e.g., mortality, function, health status, or complication) or PRO (e.g., HRQoL/function, symptom, experience, health-related behavior) (Box 1)? -> (Yes)-> Is there a relationship between the measured health outcome/PRO and at least one healthcare action (structure, process, intervention, or service) is demonstrated by empirical data (Box 2)? ->(Yes)-> PASS

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

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

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

- The developer provided the mean and distribution of Risk-Standardized Improvement Rates (RSIRs) for clinicians and clinician groups with ≥25 THA/TKA Patients with PRO data using the Full Sample Dataset which included 19,429 elective primary THA/TKA procedures from July 1, 2016 – June 30, 2018.
- The mean (SD) for Clinician-level RSIRs (Combined Dataset) and Clinician Group-level (Combined Dataset) were 64.21 percent (13.12) and 64.74 percent (12.64), respectively. The distribution of performance for the 25th to the 75th percentile ranged from 56% to 73% for the Clinician-level RSIRs (Combined Dataset) and a similar distribution for the Clinician Group-level (Combined Dataset)
- The developer notes that the mean and distribution of the RSIRs for both clinician and clinician groups from this measure supports variability in clinician and clinician group performance; therefore, there are opportunities for improving patient outcomes following elective primary THA and TKA.

Disparities

- The developer evaluated the distribution of RSIRs by quartiles of proportions of patients (n= 19, 429) with dual eligibility (n=539, 2.77 percent), low socioeconomic status (SES) using the Agency for Healthcare Research and Quality (AHRQ) SES Index (n= 1,833, 9.43 percent), and of non-white race (n= 1,483, 7.63 percent) among patients with PROs for clinicians and clinician groups with ≥ 25 THA/TKA patients with PRO data.
- Distribution of RSIRs for Clinicians and Clinician-groups (with ≥25 THA/TKA Patients with PRO data) by Proportion of Patients with Dual Eligibility with PROs
- ○ Clinicians with 0% Dual Eligible Patients among Patients with PROs
    - 25th percentile- 53.1 percent
    - 75th percentile- 72.93 percent
- ○ Clinicians with Highest Proportion of Dual Eligible Patients among Patients with PROs
    - 25th percentile- 58.66 percent
    - 75th percentile- 71.66 percent
- Distribution of RSIRs for Clinicians (with ≥25 THA/TKA Patients with PRO data) by Proportion of Patients with Low SES (AHRQ SES Index Score: Lowest Quartile) with PROs
- ○ Clinicians with Lowest Proportion of Low SES Patients among Patients with PROs
    - 25th percentile- 53.16 percent
    - 75th percentile- 72.88 percent
- ○ Clinicians with Highest Proportion of Low SES Patients among Patients with PROs
    - 25th percentile- 55.64 percent
    - 75th percentile- 77.15 percent
- Distribution of RSIRs for Clinicians (with ≥25 THA/TKA Patients with PRO data) by Proportion of Non-white Patients with PROs
- ○ Clinicians with Lowest Proportion of Non-white Patients among Patients with PROs
    - 25th percentile-50.65 percent
    - 75th percentile- 72.93 percent
- ○ Clinicians with Highest Proportion of Non-white Patients among Patients with PROs
    - 25th percentile- 57.80 percent
    - 75th percentile- 73.52 percent

Questions for the Committee:

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

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

Committee Pre-evaluation Comments:

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

1a. 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 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.

- This patient-reported structure/process measure has evidence of improvement by the use of a logic model which passed the SMP. "The evidence supports attributing patient-reported outcomes to the surgeons performing the procedure, including data supporting that low surgeon case volume is associated with longer operating times, lengthier hospitalizations, higher infection rates, and worse PROs." The evidence applies directly to the measure. The target population felt this would help them pick their future surgeon. The patient reported outcome would provide feedback to the surgeon and give other patients an idea about their proposed surgical experience.
- Yes
- Pass
- Evidence supports patient-reported outcomes directly related to the surgeons performing the procedure.
- New measure, logic model reasonable- assesses mortality/function/health status and complications as a composite
- There is a concern with the 22-point PROM improvement threshold will result in a substantial number of patients not meeting the target threshold, even if they feel better and experience no complications. The developer noted supporting evidence that attributes patient-reported outcomes to the surgeons performing the procedure, including data supporting that low surgeon case volume is associated with longer operating times, lengthier hospitalizations, higher infection rates, and worse PROs.

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?

- Yes, there was a current demonstration of a performance gap. "The mean (SD) for Clinician-level RSIRs (Combined Dataset) and Clinician Group-level (Combined Dataset) was 64.21 percent (13.12) and 64.74 percent (12.64), respectively. The distribution of performance for the 25th to the 75th percentile ranged from 56% to 73% for the Clinician-level RSIRs (Combined Dataset) and a similar distribution for the Clinician Group-level (Combined Dataset)"
- Yes, Risk-Standardized Improvement Rates were studied per Qs of studied patients.
- The distribution of RSIR supports variability in clinician performance therefore opportunities to improve patient outcomes.
- Performance data was provided. Shows gap in care and some disparities related to dual eligibility, SES, and race.
- Yes (mean and distribution of RSIRs from '16-'18 are presented); moderate disparity/gap
- Yes, there is a moderate concern for the gap in care when evaluating patients of low socioeconomic status and non-white race, although the developer did attempt to account for this discrepancy by utilizing the AHRQ index
- Data presented on entities with >=25 procedures with PROs reflect substantial room for improvement overall and important variation between entities.

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.

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

NQF Scientific Methods Panel Subgroup Evaluators: Patrick Romano, MD, MPH; Sherrie Kaplan, PhD, MPH; Daniel Deutscher, PT, PhD; Joseph Hyder, MD; John Bott, MBA, MSSW; Bijan Borah, MSc, PhD; Jack Needleman, PhD; Jennifer Perloff, PhD; Susan White, PhD, RHIA, CHDA; Ronald Walters, MD, MBA, MHA, MS; David Nerenz, Co-Chair, PhD; Sean O'Brien, PhD; Eric Weinhandl, PhD, MS

Methods Panel Evaluation Summary:

This measure was reviewed by the Scientific Methods Panel and discussed on the call. While the measure passed initial review, a member of the subgroup pulled the measure for discussion. A summary of the measure and the Panel discussion is provided below.

Reliability

- Reliability testing conducted at the Patient or Encounter level:
    - Evidence for data element reliability was provided through existing literature; the PRO-PM was originally developed for and tested at the data element level for this population.
    - The developer used test-retest and internal consistency to assess reliability of both PRO-PM instruments or PROMs (i.e., HOOS, JR and KOOS, JR). Internal consistency was calculated using the Pearson Separation Index (PSI) for both instruments. Internal consistency ranged from 0.84-0.87.
    - Intra-class correlations for reliability were between four dimensions (Pain, Symptoms, Activities of Daily Living, Sport and Recreation Function, and Quality of Life) of the HOOS, JR and the KOOS, JR with ranges from 0.75 to 0.97.
    - Reliability testing was also conducted at the Accountable Entity level:
    - The developer performed reliability testing at the measure score-level using a signal to noise ratio (SNR) approach. Among clinicians and clinician-groups with five and 10 cases, the SNR

  - yielded median reliability scores ranging from 0.70-0.79 and 0.79-0.85, respectively. The mean reliability score was 0.69 (SD 0.16) for clinicians with at least five cases.
  - Among clinicians and clinician-groups with at least 25 cases, the SNR ratio yielded median reliability scores ranging from 0.87 (mean 0.87 [SD 0.05], interquartile range [IQR] 0.09) to 0.92 (mean 0.90 [SD 0.06)], IQR 0.10), respectively.
  - One SMP member raised concern regarding variation in responses as it relates to social risk (i.e., race) and that the experiences among racial groups may be underrepresented in the sample.

Validity

- Validity testing conducted at the Patient or Encounter level:
  - The developer evaluated responsiveness for both instruments using standardized response means and then compared against two other previously validated PROMs.
    - External validity was evaluated for both instruments using Spearman's correlation.
    - a) Correlations ranged from 0.84- 0.94 for HOOS, JR testing.
    - b) Correlations ranged from 0.72- 0.91 for KOOS, JR testing.
  - The floor and ceiling effects for HOOS, JR were (0.6 percent – 1.9 percent) and (37 percent – 46 percent), respectively.
  - The floor and ceiling effects for KOOS, JR were (0.4 percent – 1.2 percent) and (18.8 percent – 21.8 percent), respectively.
  - The SMP stated concern that the 22-point PROM improvement threshold will result in a substantial percentage of patients not meeting the target threshold, even if they experience no complications and feel significantly better after surgery. A subgroup member stated that a high percentage of pre-operative patients in the sample will “fail” the measure based on very low or very high PROM scores.
  - For patients with high pre-operative PROM scores, the developer stated this is one mechanism for reducing potentially unnecessary THA or TKA surgeries that could be managed medically. The developer further added that, from the orthopedics perspective, the ceiling effect is not concerning because it encourages clinicians and clinician groups to only offer surgery to patients that have substantive symptoms so that a benefit from surgery can be seen.
- Validity testing conducted at the Accountable Entity level:
  - Face validity was assessed by asking a 17-member TEP to respond to two statements using a six-point scale.
  - Seventy-six percent either strongly or moderately agreed with the statement that this measure, as specified, will provide a valid assessment of improvement in functional status and pain following elective, primary THA/TKA. Fifty-three percent either strongly or moderately agreed with the statement that this measure, as specified, can be used to distinguish between better and worse quality care among clinicians and clinician groups.
  - The SMP noted that of the two questions asked of the TEP, only seven of TEP members strongly agreed that “the PRO-PM as specified will provide valid assessment of improvement of functional status and pain following surgery” and only three of 17 strongly agreed that “the measure can be used to distinguish between better and worse quality of care among clinicians and clinician groups.” While the majority of the TEP voted strongly agree or moderately agree, a few SMP members felt that ‘moderately agree’ was not a strong enough agreement.
  - The SMP also stated that the second question posed to the TEP, which addressed the entity level, had poorer results than the first, noting that only three TEP members strongly agreed. The SMP asked the developer to clarify why two individuals from the TEP disagreed with the

questions asked. The developer stated their findings represent the 14 TEP members who strongly agreed or moderately agreed that the measure was valid. The developer stated the two dissenting TEP members had concerns with the method for incentivizing performance and wanted to see the measure used in a broader or different data sample.
    - Some SMP members expressed interest in observing descriptive characteristics for those patients with no response to allow for construction of models to adjust for nonresponse prior to assessing reliability.
    - Several SMP members raised concern with non-response bias and the accuracy of the developer's validity assessment as 37 percent of the sample was excluded due to missing PRO scores, 10 percent due to missing risk factors, and 2 percent without clinician attribution.

- The SMP ultimately decided to not revote on the measure and passed the measure on reliability and validity with a moderate rating.

Questions for the Committee regarding reliability:

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

Questions for the Committee regarding validity:

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

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

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

Committee Pre-evaluation Comments:

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

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

- The data elements are clearly defined. There are no issues.
- No specific concerns, reliability measures are reported
- No concerns with reliability
- No concerns that the measure can be consistently implemented.
- SMP review seemed satisfactory
- There is a moderate concern for the reliability of the measure to produce consistent results.
- Specifications are fine

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

- None
- Some questions were raised regarding some non-responder bias and the accuracy of the developer's validity assessment due to missing data. However, it was felt that by the members of the SMP group that there was moderate evidence for reliability

- No concerns
- No concerns on reliability.
- internal consistency and inter-class correlations were acceptable
- As the data is grouped into 3 areas, those that responded, those that partially responded and no response, there is a degree of concern that the measure can produce the same results a high portion of the time.
- Reliability metrics were good, but calculated on complete data, not adjusted for non-response. I am concerned concern with non-response bias and the accuracy of the reliability calculations as 37 percent of the sample was excluded due to missing PRO scores, 10 percent due to missing risk factors, and 2 percent without clinician attribution

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

- Some
- No concerns
- No concerns
- No concerns on validity.
- done at patient/encounter level- reasonable correlations- but the threshold may not be reached by a high proportion of patients
- possibly (see below)
- As noted by the SMP, the face validity testing with a TEP yielded tepid endorsement for the measure. My main concern with validity relates to non-response and variation in non-response.

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?

- 2b4 External validity showed that comparisons to HOOS and KOOS showed a Spearman coefficient of > 0.8 and face validity was present in the TEP. 2b5 yes 2b6 “More than 1/3 of the testing sample (37%) was excluded due to missing PRO scores, another ~10% was excluded based on missing risk factors and ~2% were not attributable to a clinician. Therefore 50% of the admissions for CJR were excluded.”
- It does, and it has been highlighted as a potential issue.
- No concerns
- Potential threat of non-response bias. 37% excluded due to missing PRO scores, 10% due to missing risk factors, and 2% without clinician attribution.
- N/a
- There is a concern with the threat to validity with the high number of missing data and no response for the measure
- 37 percent of the sample was excluded due to missing PRO scores, 10 percent due to missing risk factors, and 2 percent without clinician attribution. Table 9 raises some questions as to whether the denominator should be limited to entities with>= 25 observations.

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?

- 2b2 This measure only excluded staged procedures that occurred in the hospital setting. Staged procedures may occur in the outpatient setting (hospital outpatient departments and ambulatory surgical setting) that would not be excluded. This may occur during the measurement period of one year. In the future, this may be an issue since more surgeries are being performed in the outpatient setting. 2b3 There is a relationship between social risk factor variables and the measure focus. The social risk factors align well with the conceptual description. It allows for feedback to the surgeon and a resource for potential future patients. The administrative database has all of the risk-adjustment variables present at the beginning of the PRO period. Yes, the risk-adjustment was tested and the C-statistic for the risk model is 0.607. This is overall acceptable and appropriate.
- The analyses indicate acceptable results.
- None identified
- No concerns with other threats to validity.
- No concerns
- Nonresponse bias for the measure could be a concern. It appears that there is ample support for risk adjustment parameters and exclusions from the measures
- Risk adjustment model and evaluation are reasonable. It is interesting to think about the risks and benefits of including the baseline PROM score in the risk adjustment model. A 20-point improvement might be different depending on where you started.

Criterion 3. Feasibility

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

- The developer indicates that the methods used to generate the data elements needed to compute the measure score can be collected by and used by healthcare personnel during the provision of care (e.g., blood pressure, lab value, diagnosis, depression score).
- Patient and/or family reported data elements may be available electronically or in paper form.
- The developer explains that most, if not all, clinical data elements can feasibly be captured in the electronic health record (EHR) as the PRO and clinical risk variable data represent standardized results that can be captured within discrete fields.
- Administrative claims data can capture prior medical history and comorbidities to augment limited clinical risk values while reducing patient and provider burden.
    - The developer recognizes the importance of electronic data capture and that not all clinicians collect data in electronic form. The measure specifications have been harmonized with electronic clinical quality process measures (eCQMs), specifically the Functional Status Outcomes for Patients Receiving Primary Total Hip Replacements and Functional Status Outcomes for Patients Receiving Primary Total Knee Replacements), that incentivize collection of the PRO data needed to calculate the measure outcome.
    - The developer reported that advancement in mobile applications and other PRO data capture forms are likely feasible to move to an electronic format.

Questions for the Committee:

- Are the required data elements routinely generated and used during care delivery?
- Are the required data elements available in electronic form (e.g., EHR or other electronic sources)?
- Is the data collection strategy ready to be put into operational use?

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

Committee Pre-evaluation Comments:

Criteria 3: Feasibility

3. Feasibility: Which of the required data elements are not routinely generated and used during care delivery? Which of the required data elements are not available in electronic form (e.g., EHR or other electronic sources)? What are your concerns about how the data collection strategy can be put into operational use?

- All of the elements are routinely obtained during routine care. The form is available in multiple languages. If necessary, paper forms are available for patient data element reporting. I have no concerns about obtaining the data.
- There are no major concerns identified.
- No concerns regarding ability to collect data electronically
- No concerns on data collection and feasibility.
- Data generated through normal provision of care; PRO are not currently routinely obtained/reported w/ THA/TKA
- Coding and data collection with the EMR support the high feasibility of the data
- The non-response numbers in Table 9 highlight the difficulty getting patients and clinicians to produce the required data elements.

Criterion 4: Usability and Use

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

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

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

Current uses of the measure

Publicly reported? ☐ **Yes** ☒ **No**

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

OR

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

Accountability program details

- The developer noted that this PRO-PM is being submitted for initial endorsement and is not currently used in any accountability program.
- The developer noted that CMS may opt to implement this measure in the Quality Payment Program (QPP) 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

- The measure is currently not implemented in a public reporting or accountability program.

Additional Feedback:

- The developer noted that they obtained input during measure development by convening a Technical Expert Panel (TEP), a Clinical Working Group, and Patient Working Group between August 2020 and July 2021.
  - The TEP was comprised of 21 total members (five of which were patients), a Clinical Working Group (four clinical expert members representing each of the four national THA and/or TKA professional societies), and a Patient Working Group (six members)
- The developer solicited feedback through teleconference meetings with the TEP (four meetings), Clinical Working Group (three meetings), and Patient Working Group (three meetings).
- The developer noted that the TEP and Clinical Working Group indicated strong support of measure specifications and provided recommendations for ongoing evaluation, such as consideration of provider volume, handling of staged procedures, the impact of social risk, and the expansion of the postoperative timeframe.
- Clinicians from the TEP and Clinical Working Group, along with the developer's clinical expert, recommended ongoing evaluation of the risk model and social risk factor analyses.
- The Patient Working Group indicated that a patient-reported, outcomes-based performance measure following elective THA or TKA procedures would be helpful for patients in selecting their surgeon, as well as supporting informed decision making.
- The hospital-level THA/TKA PRO-PM development team engaged with patients during the selection of the cohort, measure outcome, data collection instruments, and risk adjustment model.

***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**

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

4b. Usability evaluates 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

- This is a new PRO-PM, not currently used in a quality improvement program, and there are no performance results to assess.

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

- This is a new PRO-PM not yet implemented. There are no unexpected findings noted during PRO-PM development or testing by the developer.

Potential harms

- There are no harms identified by the developer.

Additional Feedback: None identified.

Questions for the Committee*:*

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

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

Committee Pre-evaluation Comments:

Criteria 4: Usability and Use

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

- The measure is new, and the intent is to eventually report it publicly via the Quality Payment Program. Yes, there has been reciprocal feedback on the measure. "TEP, Clinical Working Group, and Patient Working Group feedback has been considered throughout measure re-specification. Furthermore, the hospital-level THA/TKA PRO-PM development team engaged with patients during the selection of the cohort, measure outcome, data collection instruments, and risk adjustment model."
- Yes, it had.
- Association of periOperative Registered Nurses (AORN)
- Not currently publicly reported or part of an accountability program. CMS considering for QPP. No concerns on use.
- Due to the nature of the measure- should be useable for patient/providers, etc.
- The measure is not currently publicly reported nor used in an accountability program but will be submitted for possible inclusion in the Quality Payment Program in the foreseeable future.
- Initial measure submission so not required to be in use

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

- The measure would allow clinicians to know their patient's reported outcomes and know that various elements of their care (i.e., surgical time, low volumes) should be changed in order to deliver a higher level of healthcare. I can't perceive any untoward consequences at this time.
- No harm has been identified, and the measure was considered usable.
- No unintended consequences identified
- No unexpected findings or harms identified. No concerns on usability.
- access to date among minorities/not clear if this will be an issue to roll out in ethnic minorities where disparities frequently exist

- Yes, the results can be used to future high quality, efficient care. I do not foresee any unintended consequences in the measure and the inclusion of a patient working group in the measure development lends support for the benefit of the measure.
- potentially very usable

Criterion 5: Related and Competing Measures

Related or competing measures

- # 0425 Functional Status Change for Patients with Low Back 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)
- # 3461 Functional Status Change for Patients with Neck Impairments
- # 3493 Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) for Merit-based Incentive Payment System (MIPS) Eligible Clinicians and Eligible Clinician Groups
- # 3559 Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

Harmonization

- The developer noted the measure aligns with two non NQF-endorsed measures: Functional Status Assessment for Total Hip Replacement (QPP Quality ID: 376) and Functional Status Assessment for Total Knee Replacement (QPP Quality ID: 375). The Center for Medicare and Medicaid Services (CMS) is the steward for both measures.

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?

- They listed several NQF and non-NQF measures which they have harmonized. There does not appear to be any additional steps needed to harmonize these measures.
- it seems that the measure developer provided sufficient evidence for how this measure is harmonized/complementary with other potentially competing measures in this domain.
- Developer identified related measures #3639 is harmonized with
- Multiple related measures. No issues with harmonization.
- Several related/competing measures are listed (0422, 0425, 1550, 1551, 0424, 0423, 2643, 0426, 0428, 361, 3559, 3493, 0427)- not clear that it’s harmonized with these measures
- There are other measures that look to evaluate the risk standardized complication rate for THA and TKA as well as 30-day readmission, but they appear to have a different focus than this measure for patient reported outcomes.
- None

Public and Member Comments

Comments Submitted as of: January/17/2022

The American Medical Association (AMA)

#3639 Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM)

The American Medical Association (AMA) appreciates the opportunity to comment on NQF #3639, Clinician-Level and Clinician Group-Level Total Hip and Total Knee Arthroplasty (THA/TKA) Patient-Reported Outcome-Based Performance Measures (PRO-PMs). The AMA supports the assessment of patient-reported outcomes but believes that the burden of data collection to the clinician, practice, and patient must be adequately addressed and the continued multi-step approach to risk adjustment must be reconsidered prior to implementation of this measure in the Merit-Based Incentive Payment System (MIPS). We also request clarification on which version of this measure is undergoing endorsement review since MUC2021-107 includes a different post-operative assessment timeframe. We believe that the alignment of the timeframe with the 1-year follow-up visit as recommended by the technical expert panel feedback per MUC2021-107 is preferable.

On review of the measure specifications, we note that the information required for the numerator and risk variables includes multiple data elements from additional patient-reported surveys beyond those used to assess the patient-reported outcome of interest. Furthermore, this information is expected to be collected between 90 to 0 days prior to surgery. The AMA supports the inclusion of many of these variables within the risk model given their relevance to how patients may or may not be able to achieve improvement but questions whether CMS adequately assessed the feasibility and potential data collection burden to the clinician, practice, and patient, particularly since the data used for measure development relies on hospital reporting through the Comprehensive Care for Joint Replacement Model. The limited information on feasibility does not provide any detail on how the testing sites coordinated data collection across settings or on whom the responsibility of the additional items was placed. This question is particularly important since the specifications require clinicians and practices to collect data for one measure from 90 days pre-operatively to up to 425 days post-operatively, which the hospital is also likely collecting at the same time. The inclusion of this measure in addition to the one at the hospital-level further raises our concerns over how duplication of effort in collecting these data required for the measure numerator and risk adjustment variables can be avoided. The NQF submission form does not adequately address these concerns and the AMA urges CMS to complete additional testing around the feasibility of data collection and reduction of reporting burden prior to endorsement.

Perhaps even more importantly, we would have expected to see an assessment from the patient's perspective on whether the timing and number of items solicited throughout this process were appropriate and does not result in survey fatigue, particularly now that they may have the hospital and clinician requesting the same data. For example, would the number of surveys throughout the pre-, intra-, and post-operative timeframes lead them to be less likely to complete other surveys such as HCAHPS or CG-CAHPS? CMS should also examine if whether the timing of data collection is appropriate such as if the pre-operative PRO-PM data were collected on the morning of the surgery, could stress and anxiety have impacted responses? We believe that it is critical to understand the potential impact and burden that could be experienced. While it may seem reasonable for one measure, if this measure is an example of how future measures could be specified, what is the potential long-term impact on patients, hospitals, clinicians, and practices as more and more PRO-PMs are implemented?

The AMA also believes that measures must meet minimum acceptable thresholds of 0.7 for reliability. We urge NQF to require the developer to set the case minimum at 25 cases in order to achieve this threshold.

The AMA strongly supports the inclusion of health literacy in the risk model but remains concerned that CMS continues to test social risk factors after the assessment of clinical and demographic risk factors, and it is unclear why this multi-step approach is preferable. On review of the Evaluation of the NQF Trial period for Risk Adjustment for Social Risk Factors report, it is clear that the approaches to testing these data should be revised to strategies such as multi-level models or testing of social factors prior to clinical factors and that as access to new data becomes available, it may elucidate more differences that are unrelated to factors within a hospital's control. Additional testing that evaluates clinical and social risk factors at the same time or social prior to clinical variables rather than the current approach with clinical factors prioritized should be completed.

The AMA believes that additional information on these concerns is needed prior to endorsement of this measure. We respectfully ask the Standing Committee to consider these comments and seek additional information prior to any decision on endorsement.

Reference:

National Quality Forum. Evaluation of the NQF Trial period for Risk Adjustment for Social Risk Factors. Final report. July 18, 2017. Available at: http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=85635. Last accessed January 17, 2022.

Combined Methods Panel Scientific Acceptability Evaluation

Measure Number: *3639*

Measure Title: *Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM)*

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: Items sp.01-sp.30

   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.**

   For example: 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?

   Reviewer 2: Since data collection allows for variable modes of administration, (with 16.5% missing mode of data collection) specific protocols for handling, (e.g., follow-up for non-responders, allowing for mixed modes within practices, etc.) would have been helpful. Also, specifications regarding recording of who completed the survey (patient or proxy) are needed, as are whether the patient responded in English of Spanish to assess potential proxy or language/literacy biases.

   Reviewer 3: No concerns.

   Reviewer 5: The exclusions listed in sp. 14, are not defined as to the data source (e.g., inpatient claims) &/or specifications within the data source (e.g., specific ICD-10 codes).

Reviewer 7: None. I am accepting the developer's statements on the validation done on the individual instruments.

Reviewer 9: N/A

Reviewer 10: None.

Reviewer 11: None.

RELIABILITY: TESTING

Type of measure:

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

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

Data Source:

☒ **Claims** ☐ **eCQM (HQMF) implemented in EHRs** ☒ **Abstracted from Electronic Health Records**

☐ **Abstracted from Paper Medical Records** ☒ **Instrument-Based Data** ☐ **Registry**

☒ **Enrollment Data** ☐ **Other (please specify)**

Level of Analysis:

☒ **Group/Practice** ☒ **Individual Clinician** ☐ **Hospital/facility/agency** ☐ **Health Plan**

☐ **Population: Regional, State, Community, County or City** ☐ **Accountable Care Organization**

☐ **Integrated Delivery System** ☒ **Other (please specify)**

Reviewer 9: Medicare claims and enrollment data

Submission document: Questions 2a.01-09

3. **Reliability testing level**

   For example: for some types of measures, if patient/encounter level validity is demonstrated, additional reliability testing is not required. Please review table above.

   ☒ **Accountable-Entity Level** ☒ **Patient/Encounter Level** ☐ **Neither**

4. **Reliability testing was conducted with the data source and level of analysis indicated for this measure**

   NOTE: "level of analysis" reflects which entity is being assessed or held accountable by the measure.

   For example: If a measure is specified for a clinician level of analysis, but facility-level testing is provided, then testing does NOT match level of analysis. Or, if two levels of analysis are specified (e.g., clinician and facility) but testing is conducted for only one, then testing does NOT match level of analysis. Or, if claims data are selected as a data source, but testing data doesn't include claims data, then testing does NOT match data source.

   Also, check "NO" if only descriptive statistics are provided or submitter only describes process for data management/cleaning/computer programming.

   ☒ **Yes** ☒ **No**

5. If accountable-entity level and/or patient/encounter level reliability testing was NOT conducted or if the methods used were NOT appropriate, was **empirical VALIDITY testing** of **patient-level data** conducted?

   According to current guidance patient/encounter level validity testing can be used for patient/encounter level reliability testing. Answer ONLY if you responded "Neither" on question #3 and/or "No" to question

#4. Note that for some types of measures, additional reliability testing is not required IF patient/encounter level validity is demonstrated.

☐ **Yes** ☐ **No**

6. **Assess the method(s) used for reliability testing**

Submission document: Question 2a.10

For example: Is the method(s) appropriate? If not, please explain (and offer potential alternatives if possible). Does the testing conform to NQF criteria and guidance? Was testing was conducted with the data source and level of analysis indicated for this measure? Address each level of testing provided, and each analysis under each method.

Reviewer 2: It is not clear that the method for conducting the ICC analysis was performed correctly - i.e., the ratio of between practice variation in total scores (HOOS, JR, KOOS, JR) (numerator) and between practice variation in total scores plus within practice variation across patients within a practice plus with patients across items in the HOOS, JR/KOOS, HR). It appears that the ICC's reported for test/retest reliability were done at the patient level, averaged within a practice.

Reviewer 3: No concerns

Reviewer 4: Internal consistency and test-retest reliability ICC Score SNR.

Reviewer 5: The testing performed in regard to the measure was appropriate except for the fact that only a select number of data elements were tested (i.e., HOOS, JR & KOOS, JR) vs. all critical data elements. More specifically, the specific tests follow: [1] Data element testing: Data element testing was performed regarding: [a] HOOS, JR: Testing regarding internal consistency & test-retest reliability [b] KOOS, JR: Testing regarding internal consistency & test-retest reliability [2] Measure score testing: Testing involved signal-to-noise ratio (SNR) reliability testing of individual clinician score & group score.

Reviewer 7: Variety of methods used to assess reliability of instruments appear adequate and results satisfactory. Signal to noise use to test reliability of differentiating clinician variation. Substantial number of patients receiving treatment do not complete documentation to apply measure. This was addressed by using standardized inverse probability weights.

Reviewer 9: Literature support for data element reliability - established instruments used. Signal to noise used for measure score reliability.

Reviewer 11: Methods used were appropriate at both individual and entity levels.

Reviewer 13: Signal-to-noise ratio estimation.

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

Submission document: Question 2a.11

For example: Is the test sample adequate to generalize for widespread implementation? Is there high or moderate confidence that the measure results and/or the data used in the measure are reliable? Address each level of testing provided, and each analysis under each method.

Reviewer 2: Because of the above concerns, the data as reported are not interpretable. Additional clarity regarding the specifics of the method used is needed.

Reviewer 3: Results indicate moderate to high reliability at both data elements and entity level testing.

Reviewer 4: Median SNR 0.7 with ranges provided.

Reviewer 5: Testing results varied by test type & data source tested. Generally, findings indicate moderate to high reliability. More specifically: [1] Data element testing: Data element testing was performed regarding: [a] HOOS, JR: Internal consistency: 0.86 in the HSS cohort and 0.87 in the FORCE-TJR cohort Test-retest reliability: ICCs of 0.83 - 0.89 (Pain sub-scale), 0.86 - 0.94 (ADL sub-scale). [b] KOOS, JR:

Internal consistency: 0.84 in the HSS cohort and 0.85 in the FORCE-TJR cohort Test-retest reliability: ICCs of 0.85 (Pain sub-scale), 0.75 (ADL sub-scale), 0.93 (Symptoms). [2] Measure score testing: SNR ratio yielded a median reliability score of 0.87 (range: 0.79 – 0.97).

Reviewer 7: Instrument reliability appears acceptable. S/N meets standards applied by Methods panel. Variables used for modeling nonresponse are described, but descriptives on variations in response by variable not provided, and results of modeling (including coefficients indicating relative weight of each variable) not provided. Race was identified as a factor in nonresponse but race in the SD risk model analysis had small coefficient. I am concerned that results within racial groups might influence nonresponse and that the sample of blacks among the responders may be unrepresentative of experience of all Blacks undergoing the procedure, and upweighting them does not address this issue. (There is also an issue that non-Whites appear underrepresented in population having TKA and THA, but that is not an issue in assessing measure.)

Reviewer 9: Good results (0.79 to 0.99) for clinicians with at least n = 25 cases - lower for those with 5 or 10 cases.

Reviewer 10: Data element reliability was referred to the prior development of the HOOS and KOOS, with test-retest performed for that purpose, and reliability greater than .80 for both. Measure score reliability for this measure for clinicians with at least 25 cases was 0.87 and 0.70 for those with between five and ten cases.

Reviewer 11: Reliability for this measure is good - both at individual level and at entity level.

Reviewer 13: Signal-to-noise ratio estimates ranged from 0.7 to 0.9 across the two joints and 3 case volumes.

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: Question 2a.10-12

   For example: Appropriate signal-to-noise analysis; random split-half correlation; other accepted method with description of how it assesses reliability of the performance score.

   ☒ **Yes**

   ☐ **No**

   ☐ **Not applicable**

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

   Submission document: Question 2a.10-12

   For example: inter-abstractor agreement (ICC, Kappa); other accepted method with description of how it assesses reliability of the data elements

   Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

   ☒ **Yes**

   ☒ **No**

   ☒ **Not applicable** (patient/encounter level testing was not performed)

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

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

    ☒ **Moderate** (NOTE: Moderate is the highest eligible rating if accountable-entity 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.**

    Reviewer 2: See 6,7 above

    Reviewer 4: Mod-High.

    Reviewer 5: The response to Q2 follows: The exclusions listed in sp. 14, are not defined as to the data source (e.g., inpatient claims) &/or specifications within the data source (e.g., specific ICD-10 codes).

    Reviewer 7: The reliability based on S/N is moderate to high. Would like to see more information on the characteristics of those not providing data to allow construction of measure and models used to adjust for nonresponse before giving a rating for reliability.

    Reviewer 9: Measure should be limited to clinicians with 25 or more cases.

    Reviewer 10: The data element reliability had previously been validated for the HOOS and KOOS. Measure score reliability was high for the higher case counts, but still moderate for lower case counts.

    Reviewer 11: The analysis was well-done and results showed acceptably high reliability levels even for entities with sample sizes as small as 5.

    Reviewer 13: Excellent SNR values, even at low case volumes.

VALIDITY: TESTING

12. **Validity testing level (check all that apply):**

    ☒ **Accountable-Entity Level** ☒ **Patient or Encounter-Level** ☐ **Both**

13. **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*:* Questions 2b.01-02.

    For example: Data validity/accuracy as compared to authoritative source- sensitivity, specificity, PPV, NPV; other accepted method with description of how it assesses validity of the data elements.

    Answer NO if: only assessed percent agreement; did not assess separately for all critical data elements (or at minimum, for numerator, denominator, exclusions)

    ☒ **Yes**

    ☒ **No**

    ☐ **Not applicable** (patient/encounter level testing was not performed)

14. **Method of establishing validity at the accountable-entity level:**

    NOTE: Empirical validity testing is expected at time of maintenance review; if not possible, justification is required.

    Submission document: Questions 2b.01-02

    ☒ **Face validity**

    ☒ **Empirical validity testing at the accountable-entity level**

    ☒ **N/A (accountable-entity level testing not conducted)**

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

Submission document: Question 2b.02

For example: Correlation of the accountable-entity level on this measure and other performance measures; differences in performance scores between groups known to differ on quality; other accepted method with description of how it assesses validity of the performance score

☒ **Yes**

☒ **No**

☒ **Not applicable** (accountable-entity level testing was not performed)

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

Submission document: Question 2b.02

For example:

- *If face validity the only testing conducted: Was it accomplished through a systematic and transparent process, by identified experts, explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality, and the degree of consensus and any areas of disagreement provided/discussed?*
- *If a maintenance measure, but no empirical testing conducted, was justification provided?*
- *If construct validation conducted, was the hypothesized relationship (including strength and direction) described and does it seem reasonable?*

Reviewer 2: Construct validity and evidence of responsiveness provided in the body of the submission appear to be based on the literature and performed at the patient vs. practice levels. Table 15 of the appended report (chapter 6) suggests a substantial spread in RSIRs between practices in the physician groups with greater than or equal to 25 THA/TKA patients (mean: 64.2% (13.3), n=232), based on a multinomial logistic regression. Face validity evidence suggests that fewer TEP members thought the THA/TKA PRO-PM measures would be useful in discriminating between better and worse quality of care among clinician/clinician groups.

Reviewer 3: No concerns, except that entity level validity was not conducted, which is a requirement for a PRO-PM measure.

Reviewer 4: Face: TEP. Data element: domain correlations.

Reviewer 5: The testing performed in regard to the measure was appropriate except for the fact that: [a] only a select number of data elements were tested (i.e., HOOS, JR & KOOS, JR) vs. all critical data elements [b] very little detail provided regarding the face validity process (e.g., composition of the TEP). More specifically, the specific tests follow: [1] data element validity: [a] HOOS, JR: responsiveness & external validity [b] KOOS, JR: responsiveness & external validity [2] measure face validity: Asked a TEP & patient working group 2 questions regarding perceptions of ability of the measure to assess performance score.

Reviewer 7: Data element level validity based on original tests of instruments and correlation of scores with other related instruments. This was fine. Score level validity basically based on face validity and prior use of threshold improvements in equivalent hospital measure.

Reviewer 9: Face validity as well as external validity via comparison to registry data.

Reviewer 10: For the data element, face validity and the previous work of the HOOS and KOOS development had a Spearman coefficient of greater than 0.8. For measure score validity, a TEP was formed for face validity and correlations were calculated between the measure score and various domains.

Reviewer 11: Good empirical data from prior studies of validity of the two key outcome measures; use of TEP for face validity data at the entity level.

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

    Submission document: Questions 2b.03-04

    For example: Is the test sample adequate to generalize for widespread implementation? Do the results demonstrate sufficient validity so that conclusions about quality can be made? Do you agree that the score from this measure as specified is an indicator of quality?

    Reviewer 2: Discrimination statistics provided for lowest, highest deciles appear to reflect adequate calibration. However, the PRO submission samples indicate a relatively poor response rate (clinicians=18.5%, groups=32.3%). Further the ceiling effects or the HOOS, JR were 37% to 46%, and with the substantial clinical benefit defined at greater than or equal to 22 points for this measure, the potential for measurable improvement is of concern. This threshold (nearing 1 SD) is also above that cited in Lyman and Lee, 2018.

    Reviewer 3: No concerns for the results provided.

    Reviewer 4: The vast majority of the TEP and patients endorsed the face validity of this measure as demonstrated by the widespread agreement in responses to the two face validity statements. E.g., 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.

    Reviewer 5: Testing results varied by test type & data source tested. Generally, findings indicate moderate to high validity. More specifically: [1] data element validity: [a] HOOS, JR: Responsiveness: Response means for the HOOS, JR relative to other PROMs measuring post-surgery hip improvement were 2.38 in the HSS data and 2.03 in the FORCE registry data. External validity: Correlations tested ranged from 0.65 to 0.94. [b] KOOS, JR: Responsiveness: Response means for the KOOS, JR relative to other PROMs measuring post-surgery knee improvement were 1.79 in the HSS data and 1.70 in the FORCE registry data. External validity: Correlations tested ranged from 0.72 to 0.89. [2] measure face validity: Responses to the 1st question follow: TEP: 7 responded “Strongly Agree”, 6 responded “Moderately Agree”, and 4 responded “Somewhat Agree”. Patient work group: 2 responded “Strongly Agree” and 2 responded “Moderately Agree”. Responses to the 2nd question follow: TEP: 3 responded “Strongly Agree”, 6 responded “Moderately Agree”, 6 responded “Somewhat Agree”, and 2 responded “Somewhat Disagree”. Patient work group: 2 responded “Moderately Agree” and 2 responded “Somewhat Agree”.

    Reviewer 7: Comfortable that the measure is valid, subject to concern about how well the nonresponse bias adjustment is actually performing.

    Reviewer 9: High level of agreement with registry data.

    Reviewer 10: Yes

    Reviewer 11: Validity of the two outcomes surveys at the individual level is good. Face validity at the entity level is acceptable.

    Reviewer 13: TEP members generally felt that measures had face validity.

VALIDITY: ASSESSMENT OF THREATS TO VALIDITY

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

    Submission document: Questions 2b.15-18.

    For example: Are there exclusions? If so, are the exclusions/exceptions of sufficient frequency and variation across providers to be needed (and outweigh the data collection burden)? Are any patients or patient groups inappropriately excluded from the measure? If patient preference (e.g., informed decision-making) is a basis for exclusion, does it impact performance and if yes, is the measure specified so that the *information about patient preference and the effect on the measure is transparent? If you have concerns based on a clinical rationale, please note here as well as in question #29.*

Reviewer 2: More than 1/3 of the testing sample (37%) was excluded due to missing PRO scores, another ~10% was excluded based on missing risk factors and ~2% were not attributable to a clinician. Therefore 50% of the admissions for CJR were excluded.

Reviewer 3: No concerns.

Reviewer 5: There are exclusions listed in sp. 14 that are: a) not defined & b) not listed in the exclusions section. This section (i.e., 2b.15-18) is silent as to these exclusions in sp. 14.

Reviewer 7: Non-response bias raises issues of accuracy of assessment even with good instruments.

Reviewer 9: N/A

Reviewer 10: The exclusions are appropriate.

Reviewer 11: No concerns.

Reviewer 13: No concerns.

19. **Risk Adjustment**

Submission Document: Questions 2b.19-32

Applies to all outcome, cost, and resource use measures. Please answer all checkbox questions (19a -19d), then elaborate on your answers in your response to 19e.

19a. **Risk-adjustment method**

☐ None ☒ Statistical model ☐ Stratification

☐ Other method assessing risk factors (please specify)

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

☐ Yes ☐ No ☒ Not applicable

19c. **Social risk adjustment:**

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

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

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

19d.**Risk adjustment summary:**

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

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

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

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

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

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

For example: If measure is risk adjusted:

- *If the developer asserts there is no conceptual basis for adjusting this measure for social risk factors, do you agree with the rationale?*
- *How well do social risk factor variables that were available and analyzed align with the conceptual description provided?*
- *Are the candidate and final variables included in the risk adjustment model adequately described for the measure to be implemented?*

- *Are all of the risk adjustment variables present at the start of care (if not, do you agree with the rationale)?*
- *If social risk factors are not included in the risk-adjustment approach, do you agree with the developer's decision?*
- *Is an appropriate risk-adjustment strategy included in the measure (e.g., adequate model discrimination and calibration)?*
- *Are all statistical model specifications included, including a "clinical model only" if social risk factors are included in the final model?*

***If measure is NOT risk-adjusted**:*

- *Is a justification for not risk adjusting provided (conceptual and/or empirical)?*
- *Is there any evidence that contradicts the developer's rationale and analysis for not risk-adjusting?*

Reviewer 2: Risk adjustment with variables considered does not appear to have a significant or substantial impact on score distribution. The skew in the distribution by patient race, AHRQ SES Index and proportion of dual eligible patients in the sample may have limited the contribution of these variables to the risk model.

Reviewer 3: No concerns.

Reviewer 5: The development of the risk model is appropriate as well as the testing of the risk model. However, the findings regarding the adequacy of the risk adjustment is concerning. In response to 2b.27, the C-statistic for the risk model is 0.607. The predictive ability from lowest to highest decile is 52% to 81%.

Reviewer 7: Risk adjustment variables and rationale seems fine. Low-ish C-stat may reflect the wide variability in performance across clinicians.

Reviewer 9: Risk adjustment for demographic and clinical conditions.

Reviewer 10: The risk adjustment model utilized was developed and tested earlier for this population and involves 19 defined risk adjusters. Social risk factors utilized SES and dual eligibility.

Reviewer 11: Good conceptual discussion and empirical analyses using available data. The sample used for data analysis seemed to have a more non-white and more educated sample than would be the case nationally, so additional analyses done with broader use of the measure should be done to continuously evaluate the impact of social risk factors on the measure.

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

Submission document: Questions 2b.05-07

For cost/resource use measures, does this measure identify meaningful differences about cost and resource use between the measured entities?

Reviewer 2: It is unclear from the submission what impact if any ceiling effects and the significant clinical benefit threshold may have had on presented results, especially for the HOOS, JR measure.

Reviewer 3: No concerns.

Reviewer 5: No concerns based on the analyses conducted. However, would have preferred that an analysis identify the percent of clinicians and groups with statistically higher and lower rates.

Reviewer 7: None. Differences are substantial across clinician groups.

Reviewer 10: None.

Reviewer 11: No concerns.

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

Submission document: Questions 2b.11-14.

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***. 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.*

Note if not applicable. Note if applicable but not addressed. If multiple sets of specification (e.g., due to different data sources or methods of data collection): Do analyses indicate they produce comparable results?

Reviewer 2: Construct validity and evidence of responsiveness provided in the body of the submission appear to be based on the literature and performed at the patient vs. practice levels. Table 15 of the appended report (chapter 6) suggests a substantial spread in RSIRs between practices in the physician groups with greater than or equal to 25 THA/TKA patients (mean: 64.2% (13.3), n=232), based on a multinomial logistic regression. Face validity evidence suggests that fewer TEP members thought the THA/TKA PRO-PM measures would be useful in discriminating between better and worse quality of care among clinician/clinician groups. Discrimination statistics provided for lowest, highest deciles appear to reflect adequate calibration. However, the PRO submission samples indicate a relatively poor response rate (clinicians=18.5%, groups=32.3%). Further the ceiling effects or the HOOS, JR were 37% to 46%, and with the substantial clinical benefit defined at greater than or equal to 22 points for this measure, the potential for measurable improvement is of concern. This threshold (nearing 1 SD) is also above that cited in Lyman and Lee, 2018.

Reviewer 3: N/A

Reviewer 5: No concerns.

Reviewer 7: None.

Reviewer 9: N/A

Reviewer 10: Not applicable.

Reviewer 11: N/A

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

Submission document: Questions 2b.08-10.

For example: Are there any sources of missing data not considered? Is it clear how missing data are handled? Is missing data more of a problem for some providers or patients than others? Does the extent of missing data impact the validity of the measure?

Reviewer 3: No concerns.

Reviewer 5: No concerns.

Reviewer 7: Nonresponse is an issue. See earlier discussion. Low impact of adjustment for non-response bias on scores raises questions in my mind about success of strategy.

Reviewer 9: N/A

Reviewer 10: Reasons and proportions of missing data is provided.

Reviewer 11: The developers have carefully developed methods for dealing with missing data. Response rates in the sample were relatively low (35% or so of all possible patients), but the developers were able to use data on all potential respondents to develop adjustments for missing data.

For cost/resource use measures ONLY:

If not cost/resource use measure, please skip to question 25.

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

    Consider these specific aspects of the measure specifications: attribution, cost categories, target population.

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

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

    Attribution: Does the accountable entity have reasonable control over the costs/resources measured? Is this approach aspirational (intending to drive change) or was it developed based on current state?

    Costing Approach: Do the cost categories selected align with the measure intent, target population and care settings? Is the approach for assigning dollars to resources

    Carve Outs: Has the developer addressed how carve outs in the data source are handled (or should be handled for other users)? For example, if pharmacy data is carved out (missing) from the data set, can a measure that focuses on cost of care for asthmatics still be valid?

    Truncation (approach to outliers): What is the threshold for outliers (i.e., extremely high cost or low-cost cases) and how are they handled?

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 accountable-entity level testing has been conducted)

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

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

    ☒ **Insufficient** (NOTE: For instrument-based measures and some composite measures, testing at both the accountable-entity level and the patient/encounter 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.**

    Reviewer 3: Score level validity not tested.

    Reviewer 5: Rationale for the "low" rating based on the following: Response to Q16: The testing performed in regard to the measure was appropriate except for the fact that: [a] only a select number of data elements were tested (i.e., HOOS, JR & KOOS, JR) vs. all critical data elements [b] very little detail provided regarding the face validity process (e.g., composition of the TEP). Response to Q18: There are exclusions listed in sp. 14 that are: a) not defined & b) not listed in the exclusions section. This section (i.e., 2b.15-18) is silent as to these exclusions in sp. 14. Response to 19e: The C-statistic for the risk model is 0.607. The predictive ability from lowest to highest decile is 52% to 81%.

    Reviewer 7: While I want discussion of adjustment for nonresponse bias before approving measure, my initial assessment is that it is valid.

    Reviewer 10: Much of the rationale is carried over from the development of the original models for the HOOS and KOOS. The addition is the change in the scores on the PROM from prior to surgery to post operatively. Thus, the methodology of the previous calculation of the score and the expected change in the score is critical to the statistics and validity of the measure.

Reviewer 11: Individual-level validity of the key outcome measures is very good. Face validity for the entity level is acceptable.

Reviewer 13: Modest c-statistic of risk adjustment model.

For composite measures ONLY

If not composite, please skip this section.

Submission documents: Questions 2c.01-08

Examples of analyses:

1) If components are correlated - analyses based on shared variance (e.g., factor analysis, Cronbach's alpha, item-total correlation, mean inter-item correlation).

2) If components are not correlated - analyses demonstrating the contribution of each component to the composite score (e.g., change in a reliability statistic such as ICC, with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable, or clinical justification (e.g., correlation of the individual component measures to a common outcome measure).

3) Ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.

4) Overall frequency of missing data and distribution across providers. Ideally, sensitivity analysis of the effect of various rules for handling missing data and the rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules.

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?**

    For example: Do the component measures fit the quality construct and add value? Are the objectives of parsimony and simplicity achieved while supporting the quality construct? Do analyses demonstrate the aggregation and weighting rules fit the quality construct and rationale?

    - ☐ **High**
    - ☐ **Moderate**
    - ☐ **Low**
    - ☐ **Insufficient**

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

ADDITIONAL RECOMMENDATIONS

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

    Reviewer 10: Metric is basically quite simple - the change in a risk adjusted PROM score from before surgery to after surgery. Much of the statistical support derives from previous work with the statistics of the risk adjustment.

Developer Submission

1. 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

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Measure and Report: Evidence section. For example:

2021 Submission:

Updated evidence information here.

2018 Submission:

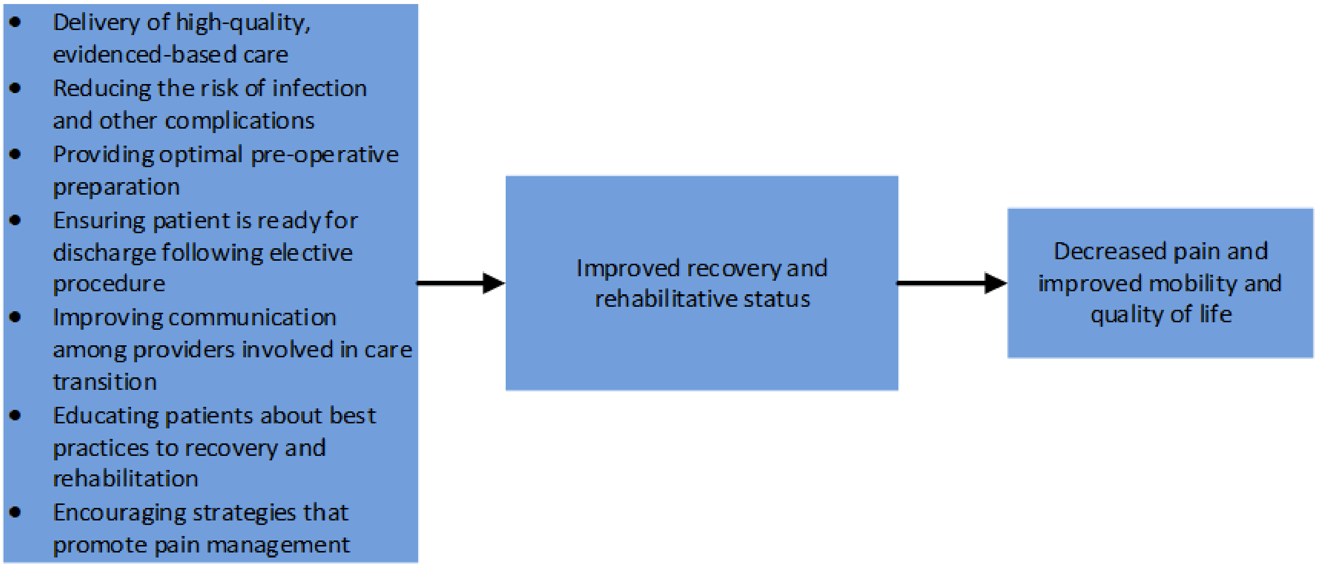
Evidence from the previous submission here.

1a. Evidence

1a.01. Provide a logic model.

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.

[Response Begins]



- Delivery of high-quality, evidenced-based care
- Reducing the risk of infection and other complications
- Providing optimal pre-operative preparation
- Ensuring patient is ready for discharge following elective procedure
- Improving communication among providers involved in care transition
- Educating patients about best practices to recovery and rehabilitation
- Encouraging strategies that promote pain management

Improved recovery and rehabilitative status

Decreased pain and improved mobility and quality of life

The goal of this measure is to directly affect patient outcomes by measuring patient-reported outcomes (PROs) 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.

[Response Ends]

1a.02. Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful.

Describe how and from whom input was obtained.

[Response Begins]

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. In alignment with the Centers for Medicare & Medicaid Services (CMS) Measures Management System (MMS), the Center for Outcomes Research & Evaluation (CORE) convened a TEP to provide feedback and recommendations on key methodological and clinical decisions in measure development. Five female patients provided input through participation in the TEP meetings in August 2020, February 2021, March 2021, and July 2021. The Patient Working Group consists of four females and two males 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 meetings in September 2020, January 2021, and June 2021.

Feedback from patients on both the TEP and the Patient Working Group indicate strong support for a clinician- and clinician group-level patient-reported outcome-based performance measure (PRO-PM) following primary elective THA and TKA. Patients were specifically enthusiastic about the development of a clinician-level PRO-PM as a tool for patients in choosing clinicians and to allow clinicians to reflect on and improve their quality of care. Patients expressed a desire to see multiple administrations of postoperative patient-reported outcome measure (PROM) surveys at select follow-up times to better capture longitudinal recovery and emphasized the importance of accounting for social determinants of health. Patients were supportive of both clinician-specific results as well as clinician group-specific measure results, noting the importance of understanding the quality of care for a clinician as well as an entire group.

During the development of the NQF-endorsed hospital-level THA/TKA PRO-PM (NQF #3559), on which this measure is based, 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. The hospital-level THA/TKA PRO-PM had significant engagement via a TEP, an Orthopedic Clinical Expert, and a Patient Working Group during development.

[Response Ends]

1a.03. Provide empirical data demonstrating the relationship between the outcome (or PRO) and at least one healthcare structure, process, intervention, or service.

[Response Begins]

Addressing quality of care for common and costly procedures such as THAs and TKAs is essential. THAs and TKAs are the most common inpatient surgeries among Medicare beneficiaries, with Medicare direct payments to hospitals for THA/TKA exceeding $15 billion annually (Miller et al., 2011). Between April 1, 2017, to October 2, 2019, there were 786,830 THA and TKA procedures performed in the inpatient setting for Medicare Fee-for-Service (FFS) beneficiaries 65 years and older (DeBuhr et al., 2021). For the US population as a whole, some project that annual THA and TKA procedures performed will reach nearly 2 million by 2030 (Lopez et al., 2020).

Complex and critical aspects of care — such as surgical approach and technique, perioperative planning, shared decision making with the patient, 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. Patient outcomes are influenced by many factors, among them patient status on presentation, and therefore this measure is adjusted to account for patient-level characteristics. Evidence supports attributing patient-reported outcomes to the surgeons performing the procedure, including data supporting that low surgeon case volume is associated with longer operating times, lengthier hospitalizations, higher infection rates, and worse PROs (Liebs et al., 2013; Lau et al., 2012; Malik et al., 2018; Levaillant et al., 2020). Additionally, in the UK, the aspect of experience most strongly associated with positive assessments of efficacy by the patient for elective surgical procedures, like THAs/TKAs, was the trust and level of communication between the patient and the surgeon, emphasizing the importance of clinician communication in shaping improvements in postoperative quality of life (Black et al., 2014).

References:

Black N, Varaganum M, Hutchings A. Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery. *BMJ Quality & Safety*. 2014; 23(7): 534.

DeBuhr J, Araas M, Grady JN, Sigler A, et. al. 2021 Procedure-Specific Complication Measure Updates and Specifications Report Hospital-Level Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) – Version 10.0. April 2021.

Lau RL, et al. The role of surgeon volume on patient outcome in total knee arthroplasty: a systematic review of the literature. *BMC Musculoskeletal Disorders*. 2012; 13(1): 250.

Levaillant M, et al. Assessing the hospital volume-outcome relationship in surgery: a scoping review protocol. *BMJ Open*. 2020; 10(10): e038201.

Liebs TR, 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(2): 239-43.

Lopez CD, Boddapati V, Neuwirth AL, Shah RP, Cooper HJ, Gellar JA. Hospital and Surgeon Medicare Reimbursement Trends for Total Joint Arthroplasty. *Arthroplasty Today.* 2020; 6(3): 437-444.

Malik AT, et al. Does Surgeon Volume Affect Outcomes Following Primary Total Hip Arthroplasty? A Systematic Review. *The Journal of Arthroplasty*. 2018; 33(10): 3329-3342.

Miller DC, Gust C, Dimick J B, et al. Large variations in Medicare payments for surgery highlight savings potential from bundled payment programs. *Health Aff*. Oct 2011, 11: 2107-15.

[Response Ends]

1b. Performance Gap

1b.01. Briefly explain the rationale for this measure.

Explain how the measure will improve the quality of care and list the benefits or improvements in quality envisioned by use of this measure.

[Response Begins]

The goal of this measure is to improve patient outcomes by providing information to patients and clinicians about clinician- and clinician group-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 surgical approach and technique, perioperative planning, shared decision making with the patient, 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 by the surgeon performing the THA or TKA procedure, but also by patient status on presentation, this measure is risk-adjusted to account for patient-level characteristics.

THA/TKA procedures provide a particularly rich test bed for developing quality measures based upon patient-reported experiences and piloting performance measures based upon PROMs. 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 a PRO-based measure via TEP and Patient Working Group engagement. Likewise, the hospital-level THA/TKA PRO-PM upon which this measure is based had strong patient support.

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

[Response Ends]

1b.02. Provide performance scores on the measure as specified (current and over time) at the specified level of analysis.

Include mean, std dev, min, max, interquartile range, and 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 (4b) under Usability and Use.

[Response Begins]

Table 1 summarizes the mean and distribution of Risk-Standardized Improvement Rates (RSIRs) for clinicians and clinician groups with ≥25 THA/TKA Patients with PRO data using the Full Sample Dataset which included 19,429 elective primary THA/TKA procedures from July 1, 2016 – June 30, 2018.

Table 1. Mean and Distribution of Risk-Standardized Improvement Rates (RSIRs) for Clinicians and Clinician Groups (with ≥25 THA/TKA Patients with PRO data) following Elective Primary THA/TKA Performed July 1, 2016 to June 30, 2018

Summary Statistics	Clinician-level RSIRs (Combined Dataset)	Clinician Group-level RSIRs (Combined Dataset)
N	232 (Clinicians)	170 (Clinician Groups)
Mean (SD)	64.21% (13.12)	64.74% (12.64)
Percentile	-	-
100% Max	88.56%	85.90%
99%	84.74%	85.42%
95%	81.81%	81.43%
90%	79.10%	79.66%
75% (Q3)	73.51%	73.49%
50% (Median)	65.75%	66.69%
25% (Q1)	56.06%	58.33%
10%	47.73%	48.52%
5%	41.40%	39.76%
1%	22.31%	21.39%
0% Min	18.36%	20.86%

Cells marked by a dash (-) are intentionally left blank.

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, 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. Include citations.

[Response Begins]

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 were not met (Ghomrawi et al., 2011; Harris et al., 2013; Jourdan et al., 2012; Suda et al., 2010). Data from this measure supports high variability in clinician and clinician group performance, as noted above.

References:

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

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.

[Response Ends]

1b.04. 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.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. 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 (4b) under Usability and Use.

[Response Begins]

Table 2 summarizes the frequency of social risk factors in the Full Sample Dataset which included 19,429 elective primary THA/TKA procedures from July 1, 2016 – June 30, 2018.

Tables 3 through 8 use the Full Sample Dataset which included 19,429 elective primary THA/TKA procedures from July 1, 2016 – June 30, 2018. To evaluate measure scores by population groups, we evaluated the distribution of risk-standardized improvement rates (RSIRs) by quartiles (or tertile) of proportions of patients with dual eligibility, low socioeconomic status (SES) using the Agency for Healthcare Research and Quality (AHRQ) SES Index, and of non-white race among patients with PROs for clinicians (Tables 3 through 5) and for clinician groups (Tables 6 through 8). These results illustrate that there are many clinicians and clinician groups that had 0% or a low proportion of patients with dual eligibility, low SES, and of non-white race among patients with PROs. Among those that care for the highest proportion of dual eligible, low SES, and non-white race patients with PROs, their performance scores are similar to those with 0% or a low proportion of patients with these factors overall.

Table 2. Frequency of Social Risk Factors among Patients in the Full Sample Dataset (Patient N = 19,429)

Variable	Frequency (%) of Social Risk Factor among Patients in the Full Sample Dataset
Total	19,429
Dual eligibility	539 (2.77%)
AHRQ SES Index: **Lowest Quartile**	1,833 (9.43%)
Race: Non-white	1,483 (7.63%)

Table 3. Distribution of RSIRs for Clinicians (with >25 THA/TKA Patients with PRO data) by Proportion of Patients with Dual Eligibility with PROs

Summary Statistics	Clinicians with 0% Dual Eligible Patients among Patients with PROs*	Clinicians with Highest Proportion of Dual Eligible Patients among Patients with PROs (≥4%-74.63%)
N (Clinicians)	117	40
Percentile	-	-
100% Max	84.97%	84.37%
99%	83.75%	84.37%
95%	79.96%	81.34%
90%	78.13%	80.22%
75% (Q3)	72.93%	71.66%
50% (Median)	66.69%	63.51%
25% (Q1)	53.16%	58.66%
10%	47.77%	50.63%
5%	40.79%	46.72%
1%	25.56%	35.74%
0% Min	21.53%	35.74%

Cells marked by a dash (-) are intentionally left blank.

*Approximately 50.4% of clinicians had no patients with dual eligibility status. Therefore, we created one category for these clinicians and then created tertiles with the remaining clinicians that had patients with dual eligibility status. The third column represented the tertile with the highest proportion of dual eligible patients among patients with PROs.

Table 4. Distribution of RSIRs for Clinicians (with >25 THA/TKA Patients with PRO data) by Proportion of Patients with Low SES (AHRQ SES Index Score: Lowest Quartile) with PROs

Summary Statistics	Clinicians with Lowest Proportion of Low SES Patients among Patients with PROs (0%-3.11%)	Clinicians with Highest Proportion of Low SES Patients among Patients with PROs (≥14.29%-47.76%)
N (Clinicians)	58	55
Percentile	-	-
100% Max	88.35%	84.97%
99%	88.35%	84.97%
95%	82.37%	84.37%
90%	78.49%	82.75%
75% (Q3)	72.88%	77.15%
50% (Median)	66.15%	66.69%
25% (Q1)	53.16%	55.64%
10%	49.97%	48.78%
5%	46.64%	46.60%
1%	33.15%	38.42%
0% Min	33.15%	38.42%

Cells marked by a dash (-) are intentionally left blank.

Table 5. Distribution of RSIRs for Clinicians (with >25 THA/TKA Patients with PRO data) by Proportion of Non-white Patients with PROs

Summary Statistics	Clinicians with Lowest Proportion of Non-white Patients among Patients with PROs (0%-3.43%)	Clinicians with Highest Proportion of Non-white Patients among Patients with PROs (≥9.68%-74.63%)
N (Clinicians)	58	55
Percentile	-	-

Summary Statistics	Clinicians with Lowest Proportion of Non-white Patients among Patients with PROs (0%-3.43%)	Clinicians with Highest Proportion of Non-white Patients among Patients with PROs (≥9.68%-74.63%)
100% Max	88.41%	84.97%
99%	88.41%	84.97%
95%	83.73%	80.20%
90%	82.37%	78.47%
75% (Q3)	72.93%	73.52%
50% (Median)	66.62%	66.11%
25% (Q1)	50.65%	57.80%%
10%	40.98%	49.01%
5%	26.33%	46.83%
1%	18.44%	35.74%
0% Min	18.44%	35.74%

Cells marked by a dash (-) are intentionally left blank.

Table 6. Distribution of RSIRs for Clinician Groups (with ≥25 THA/TKA Patients with PRO data) by Proportion of Patients with Dual Eligibility with PROs

Summary Statistics	Clinician Groups with 0% Dual Eligible Patients among Patients with PROs*	Clinician Groups with Highest Proportion of Dual Eligible Patients among Patients with PROs (≥3.70%-74.63%)
N (Clinician Groups)	64	37
Percentile	-	-
100% Max	85.34%	86.08%
99%	85.34%	86.08%
95%	80.31%	83.00%
90%	77.83%	81.05%
75% (Q3)	72.45%	75.06%
50% (Median)	65.48%	64.19%

Summary Statistics	Clinician Groups with 0% Dual Eligible Patients among Patients with PROs*	Clinician Groups with Highest Proportion of Dual Eligible Patients among Patients with PROs (≥3.70%-74.63%)
25% (Q1)	54.35%	58.95%
10%	47.74%	48.25%
5%	46.26%	38.93%
1%	21.42%	38.92%
0% Min	21.42%	38.92%

Cells marked by a dash (-) are intentionally left blank.

*Approximately 37.7% of clinician groups had no patients with dual eligibility status. Therefore, we created one category for these clinician groups and then created tertiles with the remaining clinician groups that had patients with dual eligibility status. The third column represented the tertile with the highest proportion of dual eligible patients among patients with PROs.

Table 7. Distribution of RSIRs for Clinician Groups (with ≥25 THA/TKA Patients with PRO data) by Proportion of Patients with Low SES (AHRQ SES Index Score: Lowest Quartile) with PROs

Summary Statistics	Clinician Groups with Lowest Proportion of Low SES Patients among Patients with PROs (0%-3.35%)	Clinician Groups with Highest Proportion of Low SES Patients among Patients with PROs (≥15.15%-47.76%)
N (Clinician Groups)	42	43
Percentile	-	-
100% Max	84.78%	86.08%
99%	84.78%	86.08%
95%	82.16%	80.93%
90%	77.63%	80.24%
75% (Q3)	72.29%	74.69%
50% (Median)	65.85%	69.32%
25% (Q1)	57.43%	61.36%
10%	47.89%	50.58%
5%	45.82%	47.74%

Summary Statistics	Clinician Groups with Lowest Proportion of Low SES Patients among Patients with PROs (0%-3.35%)	Clinician Groups with Highest Proportion of Low SES Patients among Patients with PROs (≥15.15%-47.76%)
1%	36.47%	37.84%
0% Min	36.47%	37.84%

Cells marked by a dash (-) are intentionally left blank.

Table 8. Distribution of RSIRs for Clinician Groups (with ≥25 THA/TKA Patients with PRO data) by Proportion of Non-white Patients with PROs

Summary Statistics	Clinician Groups with Lowest Proportion of Non-white Patients among Patients with PROs (0%-3.33%)	Clinician Groups with Highest Proportion of Non-white Patients among Patients with PROs (≥9.09%-74.63%)
N (Clinician Groups)	41	43
Percentile	-	-
100% Max	84.78%	86.08%
99%	84.78%	86.08%
95%	81.92%	80.24%
90%	79.96%	77.05%
75% (Q3)	74.69%	72.67%
50% (Median)	69.58%	65.06%
25% (Q1)	60.36%	59.18%
10%	48.25%	54.37%
5%	40.56%	49.16%
1%	21.59%	45.72%
0% Min	21.59%	45.72%

Cells marked by a dash (-) are intentionally left blank.

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, 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 above.

[Response Begins]

N/A

[Response Ends]

2. 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.

---

sp.01. Provide the measure title.

Measure titles should be concise yet convey who and what is being measured (see What Good Looks Like).

[Response Begins]

Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM)

[Response Ends]

sp.02. Provide a brief description of the measure.

Including type of score, measure focus, target population, timeframe, (e.g., Percentage of adult patients aged 18-75 years receiving one or more HbA1c tests per year).

[Response Begins]

This patient-reported outcome-based performance measure uses the same measure specifications as the NQF-endorsed (NQF # 3559) hospital-level risk-standardized improvement rate (RSIR) following elective primary THA/TKA with the following exception: this measure attributes the outcome to a clinician or clinician group. Specifically, this measure will estimate a clinician-level and/or a clinician group-level 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.

[Response Ends]

sp.04. Check all the clinical condition/topic areas that apply to your measure, below.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Surgery: General*

[Response Begins]

Musculoskeletal

Musculoskeletal: Osteoarthritis

Surgery

Surgery: Orthopedic

[Response Ends]

sp.05. Check all the non-condition specific measure domain areas that apply to your measure, below.

[Response Begins]

Care Coordination

Disparities Sensitive

Health and Functional Status

Health and Functional Status: Change

Health and Functional Status: Physical Activity

Health and Functional Status: Quality of Life

Person-and Family-Centered Care: Person-and Family-Centered Care

Safety

Safety: Complications

[Response Ends]

sp.06. Select one or more target population categories.

Select only those target populations which can be stratified in the reporting of the measure's result.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]

Elderly (Age >= 65)

[Response Ends]

sp.07. Select the levels of analysis that apply to your measure.

Check ONLY the levels of analysis for which the measure is SPECIFIED and TESTED.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*
- *Population: Population*

[Response Begins]

Clinician: Group/Practice

Clinician: Individual

[Response Ends]

sp.08. Indicate the care settings that apply to your measure.

Check ONLY the settings for which the measure is SPECIFIED and TESTED.

[Response Begins]

Inpatient/Hospital

[Response Ends]

sp.09. 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. If no URL is available, indicate “none available".

[Response Begins]

none available

[Response Ends]

sp.11. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

Attach an excel or csv file; if this poses an issue, contact staff. Provide descriptors for any codes. Use one file with multiple worksheets, if needed.

[Response Begins]

Available in attached Excel or csv file

[Response Ends]

Attachment: QPPHipKneePROPMDataDict_0729.xlsx

For the question below: state the outcome being measured. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.12. State the numerator.

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.

[Response Begins]

The numerator is the risk-standardized proportion of patients undergoing an elective primary THA or TKA who experience a 22 point or 20 point or more improvement, for hip replacement and knee replacement patients respectively between preoperative and postoperative assessments on joint-specific patient-reported outcome measures (PROMs). The patient-level improvement thresholds are an a priori, patient-defined substantial clinical benefit (SCB) threshold of improvement which is an anchor-based threshold developed using patient-report of satisfaction with change in Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS,JR)/Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS JR) scores (Lyman and Lee, 2018). This measure uses the same SCB threshold developed for the hospital-level measure, which was reviewed and recommended for endorsement by the NQF Surgery Standing Committee in 2020. 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 the hospital-level THA/TKA PRO-PM development Patient Working Group, Technical Expert Panel (TEP), Technical Advisory Group, and Orthopedic Clinical Expert.

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.

[Response Ends]

For the question below: describe how the observed outcome is identified/counted. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.13. Provide details needed to calculate the numerator.

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 sp.11.

[Response Begins]

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

Two joint-specific patient reported outcome measures (PROMs) 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)

- Patient-Reported Outcomes Measurement Information Systems (PROMIS) Global

Mental Health Score (calculated with data from the 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)

Table 1 describes each data element and if it is collected pre and/or post-operatively.

Table 1. Data Elements Collected for MIPS THA/TKA PRO-PM

Type of Element	Data Element	Collection timing
PROMs	VR-12 (all items)	Preoperative
	PROMIS-Global (all items)	Preoperative
	HOOS, JR (six items)	Pre- and postoperative
	KOOS, JR (seven items)	Pre- and postoperative
Risk Variables	SILS2 questionnaire ("How comfortable are you filling out medical forms by yourself?")	Preoperative
	BMI[a]	Preoperative
	Height[b]	Preoperative
	Weight[b]	Preoperative
	Use of Chronic (≥ 90 days) Narcotics	Preoperative
	Total Painful Joint Count: Patient-Reported Pain in Non-Operative Lower Extremity Joint ("What amount of pain have you experienced in the last week in your other knee/hip?")	Preoperative
	Quantified Spinal Pain: Patient-Reported Back Pain, Oswestry Index Question ("My BACK PAIN at the moment is")	Preoperative

[a] collection of Height and Weight together will substitute the requirement to collect BMI.

[b] collection of BMI will substitute the requirement to collect Height and Weight.

Centers for Medicare and Medicaid Services (CMS) administrative data are used to identify eligible THA/TKA procedures for the measure cohort (denominator) (ICD-10 codes for eligible THA/TKA procedures identified in the Data Dictionary accompanying this NQF submission; see Tab Cohort Inclusions) and additional risk variables, including patient demographics and clinical comorbidities (see Tab Risk Variables with PRO data and Risk Variables in Risk Modeling).

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 the hospital-level THA/TKA PRO-PM development Patient Working Group, TEP, Technical Advisory Group, and Orthopedic Clinical Expert. This measure uses the same SCB threshold developed for the hospital-level measure, which was reviewed and recommended for endorsement by the Surgery Standing Committee in 2020.

The numerator is the same as the NQF-endorsed hospital-level measure. The measure numerator was defined with extensive patient and clinician input during the development of the hospital-level THA/TKA PRO-PM. Specifically, clinical experts and patients engaged during development of the hospital-level THA/TKA PRO-PM 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. Stakeholders also strongly supported a numerator definition assessing a threshold change in PROM score over averaging patient change in PROM scores for performance measure reporting. They noted that measurement of a threshold change will distinguish patients with and without substantial clinical improvement. Comments against a reported average change included concern that a hospital whose patients all achieve average results could have a reported measure score 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 improvement, supported by patients and clinical experts, provides an easy-to-understand metric that patients found intuitive. Using a SCB threshold incentivizes providers to perform surgery on patients with greater preoperative severity and lower preoperative PROM scores, a group that might otherwise not be offered surgery, as these patients can experience substantial clinical improvement but may not reach a pre-determined postoperative state and 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. Furthermore, since the SCB was defined using published literature (Lyman and Lee, 2018) and with close input from patients and clinicians during development of the hospital-level THA/TKA PRO-PM, 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. The clinician- and clinician group-level THA/TKA PRO-PM uses the same measure outcome to align with the hospital-level THA/TKA PRO-PM and ensure usability and understanding of the measure results across settings.

NQF Staff requested clarification on issues around PROM validity; below we respond to their questions below:

NQF Question: Please clarify the following: did the developer test the accuracy and consistency of collecting data from 7 different PROMs? Are they all standardized and validated? Was the assembly of individual PROs from the PROMs tested for the assembled use? How is the data collected for each PROM

CORE response: To clarify, the measures primarily uses two procedure-specific PROMs to define the measure outcome, the HOOS,JR and KOOS,JR. Both of these PROMs are well validated surveys (Lyman et al, 2016a and Lyman et al, 2016b). The measure uses the PROMIS-Global (Hays et al., 2009) or VR-12 (Kazis et al., 2017) to assess mental health for use in the risk model. The PROMIS-Global and VR-12 are also well validated surveys. The measure also uses the SILS2 (a measure of health literacy) (Morris et al., 2006) as well as assessments of back pain (Fairbank et al., 2000) and other low extremity joint pain (Ayers et al., 2013), which are all valid patient assessments. Orthopedic surgeons and their professional societies provided specific recommendations through public comment on the initial CJR proposed rule to address concomitant low back pain and other lower extremity joint pain. These experts felt it was clinically essential to accurately capture the impact of the THA/TKA and not have the PROM scores confounded by known clinical conditions that impact knee and hip PROMs. Similarly, literacy experts and patient advocates supported the use of the SILS2 as a valid tool, citing the critical need to capture health literacy without greatly increasing patient burden. Finally, the CJR model did not specify an order of the PROs or collected risk variables to be presented to the patients nor did it ask participants to report on the order of the data collected; therefore, it was not possible to test the assembly of the PROMs.

The CJR data underwent data cleaning and quality assurance steps including, identification of missing CMS Certification Number (CCN), file conversion to comma-separated values (CSV), assessing accuracy of procedure type, patient identification, and whether each variable is the correct data type and within range, where applicable. During data cleaning and quality assurance, CORE also assessed logic such as alignment of procedure type and PROM type, identification of missing variables, and removing duplicate submissions.

The data used in measure testing was collected from hospitals voluntarily reporting PRO and risk variable data in CJR. Hospitals were allowed to choose the PRO and risk variable data collection approach and some hospitals collected data on paper, electronically, or telephone. Among submissions from performance year 4 of CJR, 49.7% were completed on paper, followed by electronic (web-based, EHR, etc) 26.7%, and telephone 7.1%. Of note, 16.5% of submission had missing mode of collection information.

NQF Question: The developer discusses SCB threshold incentives and provider practice improvements to achieve the SCB. Please add some explanation of the following considerations to your testing analysis: 1) In the era of reducing opioid use, patients may need to suffer significant pain to meet a threshold of potential PROM results increases. 2) Patients with a high pain threshold may not be considered improved candidates for potential PROM results increases, and 3) The use of potential PROM results increases may increase administrative burden of elective surgical clearances, 4) The importance to achieving the PROM results that may trigger providers "practicing to the measure". Upon full submission, please be sure to address these concerns fully in the Use section.

CORE response: Thank you for highlighting these important topics for our team's consideration.

Opioid use: Opioid use (as assessed with the variable use of Chronic [≥ 90 days] Narcotics) was evaluated as a potential risk adjustment variable during development of the hospital-level measure and was included in the final risk model based on its importance. Of note, the hospital-level THA/TKA PRO-PM which this measure is based developed the final risk model and included 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 the hospital-level THA/TKA PRO-PM measure development team's technical expert panel (TEP) and clinical experts as both clinically important and feasible. CMS will continue to monitor this issue during measure reevaluation.

High pain thresholds: The intent of THA/TKA procedures is to relieve pain and improve function, both of which are validly captured by the HOOS, JR and KOOS, JR PROMs. Further, the SCB thresholds were defined using diverse patients during development of the HOOS, JR and KOOS, JR and were then vetted again with diverse patients during measure development. Our clinical experts anticipate that the impact of high pain thresholds will not negatively impact the measure results as the PROMs ask patients to rate both their pain and functional impairment.

Burden: Collecting PROMs can increase patient and provider burden, but simultaneously helps providers focus clinical and decision-making conversations on the outcomes repeatedly shown to be the most meaningful to patients, namely pain and function.

In addition, CMS is carefully planning for potential implementation of this measure which is informed by stakeholder input and with careful consideration of clinician and clinician group burden. While patient-reported outcomes performance measures (PRO-PMs) require providers to integrate data collection into clinical workflows, this integration provides opportunity for patient reported outcomes (PROs) to inform clinical decision making and benefit patients by engaging them in discussions about potential outcomes. CMS will be mindful of the flexibility providers will need to implement the THA/TKA PRO-PM.

Unintended consequences: Thank you for sharing this concern. CMS plans to monitor for any unintended consequences of the measure.

References:

Ayers, D.C., et al., 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, 2013. 95(20): p. 1833.

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.

Fairbank, J.C. and P.B. Pynsent, The Oswestry Disability Index. Spine (Phila Pa 1976), 2000. 25(22): p. 2940-52; discussion 2952.

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. Quality of Life Research, 18(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9

Kazis, L., Rogers, W., Rothendler, J., Qian, S., Selim, A., Edelen, M., Stucky, B., Rose, A., & Butcher, E. (2017). Outcome Performance Measure Development for Persons with Multiple Chronic Conditions. In RAND Corporation. https://doi.org/10.7249/rr1844

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.

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.

Morris, N. S., MacLean, C. D., Chew, L. D., & Littenberg, B. (2006). The Single Item Literacy Screener: Evaluation of a brief instrument to identify limited reading ability. BMC Family Practice, 7(21), 1–7. https://doi.org/10.1186/1471-2296-7-21

[Response Ends]

For the question below: state the target population for the outcome. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.14. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]

The cohort (target population) includes Medicare fee-for-service (FFS) patients 65 years of age and older undergoing elective primary THA/TKA procedures.

The cohort does not include patients with hip fractures, pelvic fractures, revision THAs/TKAs, and bone metastases. The rationale for each is outlined below:

- **Facture of the pelvis or lower limbs coded in the principal or secondary discharge diagnosis fields on the index admission claim** (Note: Periprosthetic fractures must be additionally coded as POA in order to disqualify a THA/TKA from cohort inclusion, unless exempt from POA reporting.) Rationale: Patients with fractures have higher mortality, complication, and readmission rates, and the procedures are typically not elective.
- **A concurrent partial hip or knee arthroplasty procedure** Rationale: Partial arthroplasty procedures are primarily done for hip and knee fractures and are typically performed on patients who are older, frailer, and have more comorbid conditions.
- **A concurrent revision, resurfacing, or implanted device/prosthesis removal procedure** Rationale: Revision procedures may be performed at a disproportionately small number of hospitals and are associated with higher mortality, complication, and readmission rates. Resurfacing procedures are a different type of procedure involving only the joint's articular surface and are typically performed on younger, healthier patients. Elective

procedures performed on patients undergoing removal of implanted device/prostheses procedures may be more complicated.

- **Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field on the index admission claim** Rationale: Patients with these malignant neoplasms are at increased risk for complication, and the procedure may not be elective.

[Response Ends]

For the question below: describe how the target population is identified. Calculation of the risk-adjusted outcome should be described in sp.22.

sp.15. Provide details needed to calculate the denominator.

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 sp.11.

[Response Begins]

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 exactly the same as the CMS's existing measure cohort for the NQF-endorsed hospital-level THA/TKA PRO-PM.

Centers for Medicare and Medicaid Services (CMS) administrative data are 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 Cohort Inclusions.)

Please note that at this time, we do not include Medicare Advantage patients in the measure results. CMS is investigating the feasibility of including Medicare Advantage data in quality measurement. In addition, the measure does not utilize claims data after the procedure; therefore, we do not include a requirement of Part B enrollment after the procedure.

[Response Ends]

sp.16. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]

The measure has three denominator exclusions, listed below.

1. Staged Procedures

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 from the measure cohort.

2. Patients who die within 270 days of the procedure

All patients who expired within 9 months (270 days) of the THA/TKA procedure are removed from the measure cohort.

3. Patients who leave against medical advice from the inpatient index admission

Finally, patients who leave their index admission against medical advice are removed from the measure cohort.

Please note that hospice patients should not be excluded from the measure cohort because any patient undergoing a major surgery such as THA/TKA most likely has short-term survival as the primary goal.

Please also note that patients without complete PROM data, such as those that refuse to complete the PROM, are excluded from the measure results given the measure requires complete PROM data to calculate the measure outcome. Patients with incomplete or no PROM data are included in the non-response bias adjustment to alleviate potential bias. Further, CMS is exploring reporting response rate or other information along with the measure results to provide the end user of the measure results with a better sense of the sample being assessed by the measure.

Below we answer additional questions from NQF staff regarding these exclusions:

Question 1, Staged Procedures:

Please explain how staged procedures are assessed when they overlap the end and beginning of measurement periods. Is there an acceptable range in days for a staged procedure? Are all staged procedures planned? Do all staged procedures need to occur in the inpatient/acute care setting? Is it possible to have 1 inpatient and 1 outpatient surgery on the same joint? Are these procedures staged? How does that impact the denominator?

CORE response: To clarify, a “staged procedure” is a bilateral THA or TKA (both right and left hips or both right and left knees). Bilateral THAs and TKAs can be performed at the same time (these are included in the measure cohort), or during separate hospitalizations (these are the excluded “staged procedures”). Therefore, all staged procedures are planned. Theoretically, a staged procedure could be performed in different settings (for example, right THA performed inpatient followed by a left THA performed in the outpatient setting), but our clinical advisors suggest this is currently rare, although it may increase in prevalence over time.

During measure development, we only assessed staged procedures as any subsequent elective, primary THA/TKA procedure in the inpatient setting that occurred during the measurement period. In the future, we will need to assess the feasibility of extending the assessment of staged procedures to before and/or after the measurement period. Of note, this exclusion represents a small number of the total patients undergoing THA and TKA procedures in our testing dataset.

Based on discussions with our orthopedic experts, including Dr. Kevin Bozic, many staged THA/TKA procedures occur within 6 months of each other; timing is solely dependent upon provider and patient discussion of the patient’s unique situation and formal guidelines do not exist. We used the measurement period given the measure has approximately a year postoperative PRO data collection window and any procedure that occurs during the postoperative PRO data collection window may negatively impact the recovery of the first procedure and it may be challenging to distinguish the recovery for either procedure from the other when they occur within 12 months of each other. In our dataset, we found that 1,181 (91.4%) of staged procedures occurred within 1 year and 111 (8.6%) of staged procedures occurred within 2 years.

To qualify as a staged procedure in the measure, the procedure must meet the criteria of an elective primary procedure. Yes, the current cohort exclusion requires staged procedures to occur in the inpatient setting. In the future we will assess staged procedures that may occur in the outpatient setting (hospital outpatient departments and ambulatory surgical setting). In the example of 1 inpatient and 1 outpatient surgery on the same joint is unlikely a staged procedure, rather a revision or other non-elective procedure on the same joint. As noted above, this is not how we define “staged procedures”. The measure cohort does not include revision procedures in measure cohort therefore subsequent procedures on the same joint that do not meet cohort criteria would not be included in the cohort.

Question 2: AMA exclusion

Are there any other forms of AMA that are appropriate for the measure, such as patients who "fire" their providers?

At this time, we only use the discharge disposition code to identify patients who leave AMA. In the example you provide of a patient “firing” their provider, please note that this information would not be systematically captured in claims data and therefore we would be unable to investigate these instances.

[Response Ends]

sp.17. Provide details needed to calculate the denominator exclusions.

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 sp.11.

[Response Begins]

1. Staged Procedures

Patients with staged procedures in the measure period are excluded. A staged procedure is identified if a patient has more than one hospitalization for 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 “Cohort Inclusions” 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.

2. Patients who die within 270 days of the procedure

Patients who die within 270 days are unable to complete PROM data in alignment with the postoperative PROM collection timeframe. The Medicare Enrollment Database, which is updated by the Social Security Administration, is used to obtain the mortality information for Medicare beneficiaries.

3. Patients who leave against medical advice

Providers are unable to deliver full care and prepare the patient for discharge when patients leave against medical advice. Specifically, if the discharge disposition code on the index admission claim is ’7’ (Left against medical advice or discontinued care), the procedure performed during that index admission is not considered eligible for cohort inclusion.

[Response Ends]

sp.18. Provide all information required to stratify the measure results, if necessary.

Include 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 in the Data Dictionary field.

[Response Begins]

N/A; this measure is not stratified.

[Response Ends]

sp.19. Select the risk adjustment type.

Select type. Provide specifications for risk stratification and/or risk models in the Scientific Acceptability section.

[Response Begins]

Statistical risk model

[Response Ends]

sp.20. Select the most relevant type of score.

Attachment: If available, please provide a sample report.

[Response Begins]

Rate/proportion

[Response Ends]

sp.21. Select the appropriate interpretation of the measure score.

Classifies interpretation of score according to whether better quality or resource use is associated with a higher score, a lower score, a score falling within a defined interval, or a passing score

[Response Begins]

Better quality = Higher score

[Response Ends]

sp.22. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

Identify the target population; exclusions; cases meeting the target process, condition, event, or outcome; time period of data, aggregating data; risk adjustment; etc.

[Response Begins]

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 experienced more than one elective primary THA/TKA during a singular hospitalization during the measurement period, include 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.

Step 7. If patient died within 270 days of the procedure, do not include in the denominator.

Step 8. If patient was discharged against medical advice from the hospital, 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 clinician- and clinician group-level measure results are calculated by aggregating all patient-level results among patients who meet the cohort definition treated by the same clinician or clinician group.

The minimum case volume used for measure testing was 25 elective primary THA/TKA patients with complete PRO and risk variable data collected 90 – 0 days preoperatively and complete PRO data collected 270 – 365 days postoperatively. Clinician- and clinician group-specific risk-standardized improvement rates (RSIRs) are calculated as the ratio of a clinician's or clinician group'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.

[Response Ends]

sp.23. Attach a copy of the instrument (e.g., survey, tool, questionnaire, scale) used as a data source for your measure, if available.

[Response Begins]

Copy of instrument is attached.

[Response Ends]

Attachment: QPPHipKneePROPMDataDict_0729.xlsx

sp.24. Indicate the responder for your instrument.

[Response Begins]

Other (specify)

The patient is the intended respondent, but the measure allows for a caregiver to respond for the patient if the patient is unable.

[Response Ends]

sp.25. If measure is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

[Response Begins]

N/A; this PRO-PM is not based on a sample.

[Response Ends]

sp.26. Identify whether and how proxy responses are allowed.

[Response Begins]

The measure will allow for proxy responses from a caregiver and clinicians/clinician groups will report whether the PROM survey responder is the patient or a surrogate.

[Response Ends]

sp.27. Survey/Patient-reported data.

Provide instructions for data collection and guidance on minimum response rate. Specify calculation of response rates to be reported with performance measure results.

[Response Begins]

Preoperative 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) can be self-administered or collected via interview; 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 accurately 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 allows 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 better represent quality performance of clinicians and clinician groups. Physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Flexibility in rearranging clinical workflows to accommodate PRO data collection as well as accessibility of PRO data in real time can inform meaningful clinical decision making. Integration of PROs into clinician decision making can increase investment in the value of PROs in improving care and quality for clinicians and for patients, resulting in higher response rates.

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, after inclusion and exclusion criteria are applied, for which complete and matched preoperative and postoperative PRO and risk variable data have been submitted for each clinician or clinician group. 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.

[Response Ends]

sp.28. Select only the data sources for which the measure is specified.

[Response Begins]

Claims

Instrument-Based Data

Other (specify)

Medicare Enrolment Database, Master Beneficiary Summary File

[Response Ends]

sp.29. Identify the specific data source or data collection instrument.

For example, provide the name of the database, clinical registry, collection instrument, etc., and describe how data are collected.

[Response Begins]

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.

Below we provide a response to a question from NQF staff:

NQF Question: Please clarify if the use of a surrogate/interpreter for non-English speaking patients has been tested for these tools. What other tools used to calculate the measure are not available for non-English speaking patients?

CORE Response: We were unable to identify studies testing the HOOS, JR and KOOS, JR on surrogates (such as family caregivers) or use of interpreters. However, the option of completing a survey via a surrogate was provided in CJR to allow for flexibility for patients and help maximize responses. In CJR, there was no information captured on whether the patient responded to the surveys in English or another language. In discussions with patients, patients noted the importance of the role of the family caregiver in providing support, such as assisting with survey responses. In discussions with providers, many noted that when translations are not available in patients' native language, use of interpreters or family members is helpful. The full forms of the HOOS and KOOS are publicly available in several languages and work is ongoing to validate the HOOS, JR and KOOS, JR in other languages. The PROMIS-Global is translated into sixteen languages and the VR-12 is available in Spanish, Chinese and German.

[Response Ends]

sp.30. Provide the data collection instrument.

[Response Begins]

Available in attached appendix in Question 1 of the Additional Section

[Response Ends]

Measure testing must demonstrate adequate reliability and validity in order to be recommended for endorsement. Testing may be conducted for data elements and/or the computed measure score. Testing information and results should be entered in the appropriate fields in the Scientific Acceptability sections of the Measure Submission Form.

- Measures must be tested for all the data sources and levels of analyses that are specified. If there is more than one set of data specifications or more than one level of analysis, contact NQF staff about how to present all the testing information in one form.
- All required sections must be completed.
- For composites with outcome and resource use measures, Questions 2b.23-2b.37 (Risk Adjustment) also must be completed.
- If specified for multiple data sources/sets of specifications (e.g., claims and EHRs), Questions 2b.11-2b.13 also must be completed.
- An appendix for supplemental materials may be submitted (see Question 1 in the Additional section), but there is no guarantee it will be reviewed.
- Contact NQF staff with any questions. Check for resources at the Submitting Standards webpage.
- For information on the most updated guidance on how to address social risk factors variables and testing in this form refer to the release notes for the 2021 Measure Evaluation Criteria and Guidance.

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

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

2b1. Validity testing demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For instrument-based measures (including PRO-PMs) and composite performance measures, validity should be demonstrated for the computed performance score.

2b2. Exclusions are supported by the clinical evidence and are of sufficient frequency to warrant inclusion in the specifications of the measure;

AND

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately).

2b3. For outcome measures and other measures when indicated (e.g., resource use):

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 14,15 and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b4. Data analysis of computed measure scores demonstrates that methods for scoring, and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful 16 differences in performance;

OR

there is evidence of overall less-than-optimal performance.

2b5. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b6. Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.

2c. For composite performance measures, empirical analyses support the composite construction approach and demonstrate that:

2c1. the component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible; and

2c2. the aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.

(if not conducted or results not adequate, justification must be submitted and accepted)

Definitions

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

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

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

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

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

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

Please separate added or updated information from the most recent measure evaluation within each question response in the Importance to Scientific Acceptability sections. For example:

2021 Submission:

Updated testing information here.

2018 Submission:

Testing from the previous submission here.

2a. Reliability

2a.01. Select only the data sources for which the measure is tested.

[Response Begins]

Claims

Instrument-Based Data

Other (specify)

[Response Ends]

2a.02. 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).

[Response Begins]

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) model. This model provided real-world PRO data collection where participating hospitals received up to 2 points towards their overall Quality Score for successful collection of PRO data (pre-determined collection thresholds) which was used to help determine model reconciliation payments. PRO data collection began in 2016 and has been extended through December 31, 2024.

Additional data were used as follows:

Medicare Parts A and B claims data were used for identifying eligible elective primary Total Hip Arthroplasty (THA)/Total Knee Arthroplasty (TKA) procedures and for identifying patient comorbid conditions.

Medicare Part B claims for inpatient services were used to attribute patients to clinicians and clinician groups who billed for the procedure.

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 THA or TKA procedures from July 1, 2016 through June 30, 2018. 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 Patient-Reported Outcome-Based Performance Measure (PRO-PM) specifications (PRO surveys, risk variable data elements, and timing of preoperative and postoperative data collection were aligned).

Below we respond to questions from NQF staff:

NQF Question 1: Please clarify: are there data elements collected in the CJR model that are not collected in non-CJR model participants? Or are all data elements in the target population included in the CJR model? Were all patients in the CJR model included in the sample?

CORE Response: Non-CJR model participants were not required to collect the PROMs and patient or provider-reported risk variables collected in CJR (ex, health literacy). For the CJR model, participating hospitals could voluntarily submit the PROM and risk variable data. Our sample included all CJR participating hospitals that voluntarily submitted PROM and risk variable data. This data was matched to administrative claims data to assess cohort criteria, additional risk variables, and clinician attribution. We limited our final sample to procedures with complete preoperative and postoperative PROM and risk variable data that met measure cohort criteria.

NQF Question 2: Please clarify if the predetermined data collection thresholds and the added 2 points to the overall Quality Score for providers in the CJR model introduced bias to the sample. Was this tested?

CORE Response: The CJR data used for testing were incentivized at the hospital level, and it is possible response rates were impacted both by the voluntary nature of PROs and by the incremental submission thresholds for CJR PRO data over time. Given the nature of the data collection, we recommend ongoing reevaluation of the measure specifications in broader datasets over time. The CJR model did not systematically ask participants to share their rationale for voluntarily reporting PRO and risk variable data. In addition, the incentive was for data collection, not for performance on the PROMs collected, reducing the likelihood that the additional quality points provided for data collection produced biased PROM results.

[Response Ends]

2a.03. Provide the dates of the data used in testing.

Use the following format: "MM-DD-YYYY - MM-DD-YYYY"

[Response Begins]

This PRO-PM was tested on eligible procedures performed between 07-01-2016 – 06-30-2018. PRO and risk variable data were collected for patients 90 – 0 days prior to surgery and PRO data were collected 270 – 365 days following surgery. Medicare claims between 07-01-2016 – 06-30-2018 were used to identify eligible THA/TKA procedures, and Medicare claims for the 12 months prior to the procedure were used to identify a patient's comorbid conditions used in risk adjustment. Medicare Part B claims for inpatient services between 07-01-2016 – 06-30-2018 were used to attribute patients to clinicians and clinician groups who billed for the procedure. The dates for EDB, MBSF, and American Community Survey data to assess Medicare FFS status, socioeconomic indicators, and race for patients were concurrent with their procedure data.

[Response Ends]

2a.04. Select the levels of analysis for which the measure is tested.

Testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan.

Please refrain from selecting the following answer option(s). We are in the process of phasing out these answer options and request that you instead select one of the other answer options as they apply to your measure.

Please do not select:

- *Clinician: Clinician*

- *Population: Population*

[Response Begins]

Clinician: Group/Practice

Clinician: Individual

[Response Ends]

2a.05. List the measured entities 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.

[Response Begins]

For this measure, the measured entities are clinicians or clinician groups serving Medicare FFS beneficiaries aged 65 years and older. A total of 1,254 clinicians and 526 clinician groups that performed elective primary THA/TKA procedures between July 1, 2016 and June 30, 2018 were included in the dataset used for measure development and testing.

The number of measured entities (clinicians and clinician groups) varies by testing type; see Section 2.a.07 for details.

[Response Ends]

2a.06. Identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis), separated by level of analysis and data source; if a sample was used, describe how patients were selected for inclusion in the sample.

If there is a minimum case count used for testing, that minimum must be reflected in the specifications.

[Response Begins]

The number of patients varies by testing type; see Section 2a.07 for details.

[Response Ends]

2a.07. 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.

[Response Begins]

We identified 1,254 clinicians and 526 clinician groups 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 meeting cohort criteria with complete PRO data was 19,429 (**Full Sample**). These data were randomly divided 60%/40% into a **Development Dataset** and a **Validation Dataset**. There is a single **Combined Dataset** which includes 232 clinicians and 170 clinician groups from the full sample dataset with at least 25 THA/TKA patients with PRO data during the measurement period.

Table 1 and Table 2 include a summary of the number of clinicians and clinician groups, respectively in each dataset as well as the mean % of patients on Medicaid and mean percentage of patients in the lowest quartile of the AHRQ SES index.

Development Dataset: Of the 11,653 patients included in this dataset, 4,193 had a THA procedure and 7,460 had a TKA procedure. Characteristics of the 11,653 patients in the dataset are presented in Table 3.

Characteristics		Development Dataset, N (%)	Validation Dataset, N (%)
Total N		11,653	7,776
Age, Mean (SD)		73.73 (5.72)	73.70 (5.74)
Male		4,405 (37.80%)	2,889 (37.15%)
BMI, Mean (SD)		30.21 (5.93)	30.35 (6.01)
Index admissions with an elective THA procedure		4,193 (35.98%)	2,778 (35.73%)
Index admissions with an elective TKA procedure		7,460 (64.02%)	4,998 (64.27%)
Number of procedures (two vs. one)		67 (0.57%)	49 (0.63%)
Mental Health Score, Mean (SD)		50.03 (8.11)	49.96 (8.09)
Health Literacy	Not at all	2,015 (17.29%)	1,267 (16.29%)
	A little bit	881 (7.56%)	621 (7.99%)
	Somewhat	1,291 (11.08%)	833 (10.71%)
	Quite a bit	2,079 (17.84%)	1,410 (18.13%)
	Extremely	5,387 (46.23%)	3,645 (46.88%)
Other Joint Pain	None	4,057 (34.82%)	2,637 (33.91%)
	Mild	2,897 (24.86%)	1,871 (24.06%)
	Moderate	2,890 (24.80%)	2,007 (25.81%)
	Severe	1,470 (12.61%)	1,046 (13.45%)
	Extreme	339 (2.91%)	215 (2.76%)
Back Pain	None	4,459 (38.26%)	2,869 (36.90%)
	Very Mild	2,905 (24.93%)	1,979 (25.45%)
	Moderate	2,964 (25.44%)	2,024 (26.03%)
	Fairly Severe	948 (8.14%)	653 (8.40%)
	Very or Worst Severe	377 (3.24%)	251 (3.23%)
Use of Chronic (≥90 days) Narcotics		2,032 (17.44%)	1,358 (17.46%)
Severe infection; other infectious diseases (CC 1, 3–7)		2,023 (17.36%)	1,386 (17.82%)

Characteristics		Development Dataset, N (%)	Validation Dataset, N (%)
Liver disease (CC 27–31)		491 (4.21%)	322 (4.14%)
Diabetes mellitus (DM) or DM complications (CC 17–19, 122–123)		3,013 (25.86%)	2,005 (25.78%)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease (CC 40)		1,249 (10.72%)	834 (10.73%)
Depression (CC 61)		1,832 (15.72%)	1,180 (15.17%)
Other Psychiatric Disorders (CC 63)		1,839 (15.78%)	1,260 (16.20%)
Coronary atherosclerosis or angina (CC 88–89)		2,878 (24.70%)	1,872 (24.07%)
Vascular or circulatory disease (CC 106–109)		2,256 (19.36%)	1,471 (18.92%)
Renal failure (CC 135–140)		1,637 (14.05%)	1,116 (14.35%)
Dual Eligibility		315 (2.70%)	224 (2.88%)
Low SES: AHRQ SES Index lowest quartile*		1,146 (9.83%)	687 (8.83%)
Race	Unknown	190 (1.63%)	125 (1.61%)
	White	10,760 (92.34%)	7,186 (92.41%)
	Black	408 (3.50%)	273 (3.51%)
	Other	113 (0.97%)	76 (0.98%)
	Asian	81 (0.70%)	56 (0.72%)
	Hispanic	64 (0.55%)	37 (0.48%)
	North American Native	37 (0.32%)	23 (0.30%)

Validation Dataset: Of the 7,776 patients included in this dataset, 2,778 had a THA procedure and 4,998 had a TKA procedure. Characteristics of the 7,776 patients in the dataset are presented in Table 3.

Combined Dataset:

This dataset includes 232 clinicians and 170 clinician groups from the total dataset with at least 25 THA/TKA patients with PRO data during the measurement period. Table 4 shows the distribution of patient volumes for Clinicians and Clinician Groups in the

Combined Dataset

(includes all patients for clinicians or clinician groups with at least 25 patients).

Table 1. Characteristics of Clinicians in Development and Validation Datasets and Full Sample

Characteristics	Clinicians in Development Dataset	Clinicians in Validation Dataset	Clinicians in Full Sample
Total Clinicians, N	1,144	1,021	1,254
Mean % of Patients on Medicaid (SD)	4.45% (15.79%)	4.56% (15.69%)	4.86% (15.13%)
Mean % of patients with low AHRQ SES Index Score	10.75% (20.35%)	9.76% (20.91%)	10.35% (18.79%)

Table 2. Characteristics of Clinician Groups in Development and Validation Datasets and Full Sample

Characteristics	Clinician Groups in Development Dataset	Clinician Groups in Validation Dataset	Clinician Groups in Full Sample
Total Clinician Groups, N	484	448	526
Mean % of Patients on Medicaid (SD)	5.75% (17.61%)	5.82% (16.05%)	6.48% (17.21%)
Mean % of patients with low AHRQ SES Index Score	10.68% (18.63%)	10.03% (19.70%)	10.45% (17.58%)

Table 3. Patient Characteristics in Development and Validation Datasets

*Note: Missing AHRQ SES Index information in Development Dataset=29 (0.25%) and Validation Dataset=12 (0.15%)

Table 4. Distribution of Volumes for Clinicians and Clinician Groups with ≥25 THA/TKA Patients with Complete PRO Data (July 1, 2016 – June 30, 2018)

Characteristic	Eligible clinicians	Eligible clinician groups
Number of entities	232	170
Median (interquartile range) number of admissions per entity	43 (30-72)	71 (38–135)
Range (min. – max.) number of admissions per entity	25–188	25–476

[Response Ends]

2a.08. List 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.

[Response Begins]

Social Risk Factors (SRFs) 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, the hospital-level THA/TKA PRO-PM 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 final risk model; we therefore do not include health literacy in the specific social risk factor testing.

References:

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.

[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.07 check patient or encounter-level data; in 2a.08 enter “see validity testing section of data elements”; and enter “N/A” for 2a.09 and 2a.10.

2a.09. Select the level of reliability testing conducted.

Choose one or both levels.

[Response Begins]

Patient or Encounter-Level (e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements)

Accountable Entity Level (e.g., signal-to-noise analysis)

[Response Ends]

2a.10. For each level of reliability testing 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.

[Response Begins]

Data Element Reliability

Data element reliability is evidenced by reliability testing conducted during the development and validation of the joint-specific 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 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 using the PSI. The PSI was used in two data samples, the HSS cohort and the 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. 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 clinician performance, the measured entity is the clinician or clinician group, and reliability is the extent to which repeated measurements of the same clinician or clinician group give similar results. We identified the clinicians and clinician groups with at least 5, 10, and 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 clinician or clinician group 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.

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.

[Response Ends]

2a.11. For each level of reliability testing checked above, what were the statistical results from reliability testing?

For example, provide the percent agreement and kappa for the critical data elements, or distribution of reliability statistics from a signal-to-noise analysis. For score-level reliability testing, when using a signal-to-noise analysis, more than just one overall statistic should be reported (i.e., to demonstrate variation in reliability across providers). If a particular method yields only one statistic, this should be explained. In addition, reporting of results stratified by sample size is preferred (pg. 18, NQF Measure Evaluation Criteria).

[Response Begins]

Data Element Reliability Results

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 PSI. Internal consistency of the HOOS, JR on the PSI were 0.86 in the HSS cohort and 0.87 in the 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 HOOS in several validation studies (Klassbo et al, 2003; de Groot et al, 2007; Ornetti et al, 2010; Nilsdotter & Bremander, 2011). 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 using the PSI. Internal consistency of the KOOS, JR on the PSI were 0.84 in the HSS cohort and 0.85 in the 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 KOOS (Roos et al, 1998). 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

For clinicians with at least 25 cases, the signal-to-noise ratio yielded a median reliability score of 0.87 (range: 0.79 – 0.97). Interquartile range was 0.09. For clinician groups with at least 25 cases, the signal-to-noise ratio yielded a median reliability score of 0.92 (range: 0.79 – 0.99). Interquartile range was 0.10. See Table 5 below for further detail.

For clinicians and clinician groups with at least 5 and 10 cases, the signal-to-noise ratio yeilded median reliability socres at or above 0.70.

Table 5. Signal to Noise Reliability, Clinicians and Clinician Groups

Characteristics	N	Median	Mean (SD)	Min	Max	Interquartile Range		
						Q1	Q3	Range
Clinicians with Volume ≥5 THA/TKA Procedures with Complete PRO Data	716	0.70	0.69 (0.16)	0.44	0.97	0.55	0.82	0.26
Clinicians with Volume ≥10 THA/TKA Procedures with Complete PRO Data	469	0.79	0.78 (0.10)	0.61	0.97	0.87	0.79	0.17
Clinicians with Volume ≥25 THA/TKA Procedures with Complete PRO Data	232	0.87	0.87 (0.05)	0.79	0.97	0.82	0.92	0.09
Clinician Groups with Volume ≥5 THA/TKA Procedures with Complete PRO Data	348	0.79	0.75 (0.17)	0.43	0.99	0.60	0.91	0.31
Clinician Groups with Volume ≥10 THA/TKA Procedures with Complete PRO Data	268	0.85	0.83 (0.11)	0.60	0.99	0.74	0.93	0.19
Clinician Groups with Volume ≥25 THA/TKA Procedures with Complete PRO Data	170	0.92	0.90 (0.06)	0.79	0.99	0.85	0.95	0.10

Cell left intentionally blank.

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.

[Response Ends]

2a.12. Interpret the results, in terms of how they demonstrate reliability.

(In other words, what do the results mean and what are the norms for the test conducted?)

[Response Begins]

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) indicate 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 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 median signal-to-noise reliability scores of 0.87 and 0.92 for clinicians and clinician groups (with at least 25 cases), respectively, indicate excellent reliability. At the threshold of at least 5 and 10 cases, the median reliability scores were all at or above 0.7 indicating acceptable reliability.

References:

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.

[Response Ends]

2b. Validity

2b.01. Select the level of validity testing that was conducted.

[Response Begins]

Patient or Encounter-Level (data element validity must address ALL critical data elements)

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)

[Response Ends]

2b.02. 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.

[Response Begins]

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 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).

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.

Measure Face Validity

We assessed face validity by asking Technical Expert Panel (TEP) and patient working group members to rate the measure according to the following two statements using a six-point scale (1 = Strongly Agree, 2 = Moderately Agree, 3 = Somewhat Agree, 4 = Somewhat Disagree, 5 = Moderately Disagree, 6 = Strongly Disagree):

- Question #1: The clinician- and clinician group-level THA/TKA PRO-PM as specified will provide a valid assessment of improvement in functional status and pain following elective, primary THA/TKA.
- Question #2: The clinician- and clinician group-level THA/TKA PRO-PM as specified can be used to distinguish between better and worse quality care among clinicians and clinician groups.

[Response Ends]

2b.03. Provide the statistical results from validity testing.

Examples may include correlations or t-test results.

[Response Begins]

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).

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.

Face Validity Results

Question #1:

Among the 17 TEP members who provided responses, 7 responded “Strongly Agree”, 6 responded “Moderately Agree”, and 4 responded “Somewhat Agree” to this question.

Among the 4 Patient Working Group members who provided responses, 2 responded “Strongly Agree” and 2 responded “Moderately Agree” to this question.

Question #2:

Among the 17 TEP members who provided responses, 3 responded “Strongly Agree”, 6 responded “Moderately Agree”, 6 responded “Somewhat Agree”, and 2 responded “Somewhat Disagree” to this question.

Among the 4 Patient Working Group members who provided responses, 2 responded “Moderately Agree” and 2 responded “Somewhat Agree” to this question.

[Response Ends]

2b.04. Provide 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?)

[Response Begins]

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).

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.

Face Validity

The vast majority of the TEP and patients endorsed the face validity of this measure as demonstrated by the widespread agreement in responses to the two face validity statements.

[Response Ends]

2b.05. 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 in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]

Meaningful differences in performance measure scores are assessed by calculating the distribution of clinician- and clinical group-level RSIRs. Variation in clinician- and clinician group-level RSIRs indicate a clinically meaningful quality gap in the delivery of care to patients undergoing elective primary THA/TKA, as some clinicians and clinician groups can achieve substantially higher rates than the average performer, while other clinicians and clinician groups perform much worse than an average performer.

In addition, statistically significant differences were assessed using a median odds ratio (MOR) (Merlo et al, 2006). The MOR represents the median increase in odds of the patient outcome (a substantial clinical benefit [SCB] improvement in PROM score from preoperative to postoperative assessment) if a procedure on a single patient was performed by a higher performing clinician or clinician group compared to a lower performing clinician or clinician group. It is calculated by taking all possible combinations of clinicians and clinician groups. Always comparing the higher performing clinicians and clinician groups to the lower performing clinicians and clinician groups. 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.

[Response Ends]

2b.06. Describe the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities.

Examples may include 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.

[Response Begins]

Table 6 and Table 7 provide the mean and distribution of clinicians' and clinician groups' RSIRs, respectively. Clinician RSIRs ranged from 18.36% to 88.56% (median: 65.75%). Clinician group RSIRs ranged from 20.86%-85.90% (median: 66.69%).

Results of the analyses to examine the MOR was 1.95 for clinicians, with upper and lower 95% confidence bands of 1.85 and 2.06. The MOR was 1.94 for clinician groups, with upper and lower 95% confidence bands of 1.80 and 2.07.

Table 6. Mean and Distribution of RSIRs for Risk Model of SCB Improvement following Elective Primary THA/TKA (Clinicians with >25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	RSIRs (Combined Dataset)
N (Clinicians)	232 (Clinicians)
Mean (SD)	64.21% (13.12)
Percentile	-
100% Max	88.56
99%	84.74
95%	81.81
90%	79.10
75% (Q3)	73.51
50% (Median)	65.75
25% (Q1)	56.06
10%	47.73
5%	41.40
1%	22.31
0% Min	18.36

Cells marked by a dash (-) are intentionally left blank.

Table 7. Mean and Distribution of RSIRs for Risk Model of SCB Improvement following Elective Primary THA/TKA (Clinician Groups with >25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	RSIRs (Combined Dataset)
N (Clinician Groups)	170
Mean (SD)	64.74% (12.64%)

Summary Statistics	RSIRs (Combined Dataset)
Percentile	-
100% Max	85.90%
99%	85.42%
95%	81.43%
90%	79.66%
75% (Q3)	73.49%
50% (Median)	66.69%
25% (Q1)	58.33%
10%	48.52%
5%	39.76%
1%	21.39%
0% Min	20.86%

Cells marked by a dash (-) are intentionally left blank.

[Response Ends]

2b.07. Provide 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.

In other words, what do the results mean in terms of statistical and meaningful differences?

[Response Begins]

The variation in RSIRs for clinicians and clinician groups (Tables 6 and 7) suggests that there are meaningful differences in performance measure scores across clinicians and clinician groups. The interquartile range represents a difference of 17.45 percentage points for clinician RSIRs and 15.16 percentage points for clinician groups, and the difference between the 10th and 90th percentiles (47.73% and 79.10% for clinicians and 48.52% and 79.66% for clinician groups, respectively) is 31.37 percentage points for clinicians and 31.14 percentage points for clinician groups. This variation indicates an important quality gap among clinicians and clinician groups.

These MORs suggest almost a 1.95-fold and 1.94-fold increase in the odds of SCB improvement by higher performing clinicians and clinician groups compared to lower performing clinicians and clinician groups. The MOR values indicate that a patient is 1.95 times greater odds to achieve SCB improvement if their elective primary THA/TKA procedure was performed by a higher performing clinician and 1.94 times greater odds if performed by a higher performing clinician group than by a lower performing clinician or clinician group.

[Response Ends]

2b.08. Describe the method of testing conducted to identify the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders). Include how the specified handling of missing data minimizes bias.

Describe the steps—do not just name a method; what statistical analysis was used.

[Response Begins]

Due to the voluntary nature of PRO survey data, we understand that accounting for potential non-response bias is important for this measure. The Center for Outcomes Research and Evaluation (CORE)'s hospital-level THA/TKA PRO-PM team conducted a thorough literature search and 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, Harvard Medical School and Harvard School of Public Health), the hospital-level THA/TKA PRO-PM team 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). We have applied the same approach to development of the current measure.

For this approach, we performed the following steps:

1. All eligible THA/TKA procedures performed within the 238 CJR participating hospitals submitting complete PRO data during the measurement period among 1254 clinicians and 526 clinician groups submitting complete PRO and risk variable data for at least one of these procedures were identified via CMS claims data (N=77,661 procedures).
2. These eligible THA/TKA procedures were categorized into one of three PRO response groups:
    - Procedures for which complete PRO and risk variable preoperative data and complete PRO postoperative data were submitted ("complete PRO submission," N=19,429).
    - 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=17,220).
    - Procedures for which no PRO data were submitted ("no response," N=41,012).
3. We compared patient characteristics and clinical comorbidities across the three PRO response groups and determined there were statistical differences in case-mix.
4. The hospital-level THA/TKA PRO-PM team 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 SES, and postoperative complication following hip or knee procedures (Hutchings et al, 2012; de Rooij et al, 2018); Patel et al, 2015; Schamber et al, 2013). These variables were included in the multinomial logistic regression.
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 weighting (IPW) were calculated for each of the three groups. For the complete responders, the stabilized weights were calculated using the following formula: where 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: Table 8 provides the distribution of the stabilized weights with mean 1.00 and standard deviation of 0.25.
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 1,254 clinicians and 526 clinician groups submitting at least one complete PRO submission for an eligible THA/TKA procedure during the measurement period, 713 (0.91%) 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.

Table 8. Distribution of Stabilized Weights Applied to Patients with Complete PRO Submission (Responders)

Summary Statistics	Stabilized Weights
Mean (SD)	1.00 (0.25)
Percentile	-
100% Max	4.96
99%	1.65
95%	1.28
90%	1.14
75% (Q3)	1.03
50% (Median)	0.96
25% (Q1)	0.90
10%	0.85
5%	0.82
1%	0.76
0% Min	0.57

Cells marked by a dash (-) are intentionally left blank.

Below we respond to a question from NQF staff:

NQF Question: Were any of the 713 deaths related to complication from the THA/TKA surgeries? Was this assessed?

CORE Response: In our testing dataset, we assessed the proportion of patients who experienced an in-hospital death versus death which occurred after discharge from their procedure and within 9 months of their procedure. We find a small proportion of patients have in-hospital deaths which are likely related to complications from the THA/TKA procedure. Specifically, among procedures with PRO data who passed away within 9 months of their procedure (n=300); n=12 (4%) had an in-hospital death and n=288 (96%) passed away after their procedures. Among all the procedures (including procedures with complete, incomplete, and no PROM and risk variable data) (N=713); n=33 (4.6%) had an in-hospital death and n=680 (95.4%) passed away after their procedure. It is challenging to pinpoint the cause of death for deaths that occur after discharge from their THA/TKA procedure.

In addition, while we exclude patients without complete pre- and post-operative PROM and risk variable data, these patients and their deaths are captured in the harmonized, NQF endorsed NQF# MIPS THA/TKA complication measure.

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.

[Response Ends]

2b.09. Provide the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data.

For example, provide results of sensitivity analysis of the effect of various rules for missing data/non-response. If no empirical sensitivity analysis was conducted, identify the approaches for handling missing data that were considered and benefits and drawbacks of each).

[Response Begins]

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 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 meeting cohort criteria performed during the measurement period by all the clinicians and clinician groups in the dataset for which complete and matched preoperative and postoperative PRO and risk variable data were submitted. With this operational definition, the mean response rate across clinicians was 32.23% (SD 24.55%) and 31.85% (SD 24.20%) across clinician groups. Among clinicians with >25 elective primary THA/TKA patients with PRO data during the measurement period, the mean response rate was 42.09% (SD 16.98%); among clinician groups with >25 elective primary THA/TKA patients with PRO data during the measurement period, the mean response rate was 36.65% (SD 18.38%) (see Tables 9 and 10, below).

Response rates may have been impacted by hospital submission thresholds set by CJR. The CJR model within which these PRO data were collected, required that hospitals submitting the data meet either a minimum percentage or an absolute minimum number of PRO cases to qualify for the quality point incentive; the thresholds in CJR performance years one, two, three, and four were 50% of or 50 eligible cases; 60% of or 75 eligible cases; 70% or 100 eligible cases; and ≥ 80% or ≥ 200 eligible procedures, respectively.

To address potential response bias, we used stabilized inverse probability weighting, created with a multinomial logistic regression to calculate stabilized inverse probability weights.

Results of the stabilized inverse probability weighting to address potential non-response bias are reflected in the comparison of mean and distribution of clinician and clinician group RSIRs for risk-adjusted model of SCB improvement with and without stabilized inverse probability weighting (Tables 11 and 12, below).

Table 9. Mean and Distribution of Clinician Response Rates (for Complete PRO and Risk Variable Data)

Summary Statistics	PRO Submission Rates (All Clinicians)	PRO Submission Rates (Clinicians with ≥25 THA/TKA Patients with PRO Data)
N (Clinicians)	1,254	232
Mean (SD)	32.23% (24.55%)	42.09% (16.98%)
100% Max	100.00%	89.47%
99%	100.00%	85.44%
95%	77.78%	70.73%
90%	66.67%	62.26%
75% Q3	50.00%	54.30%
50% Median	27.07%	41.82%
25% Q1	11.43%	27.84%
10%	4.31%	19.61%
5%	2.50%	16.83%
1%	0.96%	12.25%
0% Min	0.31%	10.98%

Table 10. Mean and Distribution of Clinician Response Rates (for Complete PRO and Risk Variable Data)

Summary Statistics	PRO Submission Rates (All Clinician Groups)	PRO Submission Rates (Clinician Groups with ≥25 THA/TKA Patients with PRO Data)
N (Clinician Groups)	**526**	**170**
Mean (SD)	**31.85% (24.20%)**	**36.65% (18.38%)**
100% Max	**100.00%**	**84.48%**

Summary Statistics	PRO Submission Rates (All Clinician Groups)	PRO Submission Rates (Clinician Groups with ≥25 THA/TKA Patients with PRO Data)
99%	100.00%	84.44%
95%	77.78%	68.18%
90%	65.38%	60.25%
75% (Q3)	48.91%	50.54%
50% (Median)	27.40%	36.31%
25% (Q1)	11.11%	21.33%
10%	4.31%	13.70%
5%	2.64%	8.77%
1%	0.96%	4.35%
0% Min	0.31%	2.89%

Table 11. Mean and Distribution of Clinician RSIRs With and Without Stabilized Inverse Probability Weighting for Potential Non-Response Bias (Combined Dataset, Clinicians with ≥25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	Risk-Standardized Improvement Rates (No Weighting)	Risk-Standardized Improvement Rates (Weighted for Non-Response)
N (Clinicians)	232	232
Mean (SD)	64.21% (13.12%)	64.09% (13.18%)
Percentile	-	-
100% Max	88.56%	88.41%
99%	84.74%	85.80%
95%	81.81%	82.37%
90%	79.10%	79.53%
75% (Q3)	73.51%	73.44%
50% (Median)	65.75%	66.10%
25% (Q1)	56.06%	55.95%
10%	47.73%	47.77%

Summary Statistics	Risk-Standardized Improvement Rates (No Weighting)	Risk-Standardized Improvement Rates (Weighted for Non-Response)
5%	41.40%	40.98%
1%	22.31%	22.33%
0% Min	18.36%	18.45%

Cells marked by a dash (-) are intentionally left blank.

Table 12. Mean and Distribution of Clinician Group RSIRs With and Without Stabilized Inverse Probability Weighting for Potential Non-Response Bias (Combined Dataset, Clinician Groups with ≥25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	Risk-Standardized Improvement Rates (No Weighting)	Risk-Standardized Improvement Rates (Weighted for Non-Response)
N (Clinician Groups)	170	170
Mean (SD)	64.74% (12.64%)	64.59 (12.77%)
Percentile	-	-
100% Max	85.90%	86.08%
99%	85.42%	85.34%
95%	81.43%	81.30%
90%	79.66%	79.74%
75% (Q3)	73.49%	73.24%
50% (Median)	66.69%	66.57%
25% (Q1)	58.33%	57.43%
10%	48.52%	46.67%
5%	39.76%	39.06%
1%	21.39%	21.59%
0% Min	20.86%	21.42%

Cells marked by a dash (-) are intentionally left blank.

[Response Ends]

2b.10. Provide your interpretation of the results, in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and non-responders), and how the specified handling of missing data minimizes bias.

In other words, 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 was conducted, justify the selected approach for missing data.

[Response Begins]

We assessed the non-response bias by the Pearson correlation between the Pearson residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. This correlation among clinicians was -0.00784 (p-value=0.27) and among clinician groups was -0.00709 (p-value=0.32). This indicates that there is not a significant association between the residuals and the probability of response.

The correlation between RSIR unadjusted and inverse probability weighted RSIR is very high (0.9958 for clinicians and 0.9956 for clinician groups) 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, in alignment with the hospital-level THA/TKA PRO-PM, and empirically in our data.

The comparison of clinician and clinician group RSIRs for risk-adjusted model of SCB improvement with stabilized inverse probability weighting and without stabilized inverse probability weighting 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 SES and health status. We therefore retained response bias adjustment for the measure results.

[Response Ends]

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 eCQMs). 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.

2b.11. Indicate whether there is more than one set of specifications for this measure.

[Response Begins]

No, there is only one set of specifications for this measure

[Response Ends]

2b.12. 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. Indicate what statistical analysis was used.

[Response Begins]

[Response Ends]

2b.13. Provide the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications.

Examples may include correlation, and/or rank order.

[Response Begins]

[Response Ends]

2b.14. Provide your interpretation of the results in terms of the differences in performance measure scores for the same entities across the different data sources/specifications.

In other words, what do the results mean and what are the norms for the test conducted.

[Response Begins]

[Response Ends]

2b.15. Indicate whether the measure uses exclusions.

[Response Begins]

Yes, the measure uses exclusions.

[Response Ends]

2b.16. Describe the method of testing exclusions and what was tested.

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?

[Response Begins]

We include the number and percentages of patients removed from the cohort for each of the three measure cohort exclusions. We also assessed the proportions of staged procedures excluded from the analysis when we considered all procedures performed by clinicians in our dataset (regardless of complete PRO data).

[Response Ends]

2b.17. Provide 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.

[Response Begins]

Among the included THA/TKA procedures, 103 (0.01%) procedures of patients who left against medical advice were excluded.

Among the included THA/TKA procedures that submitted PRO data, 2,704 (5.73%) staged procedures during the measurement period were excluded. When we consider all procedures performed by clinicians in our dataset (regardless of complete PRO status), the percentage of staged procedures excluded is 3.34%.

Among the included THA/TKA procedures that submitted PRO data, 300 (0.64%) procedures of patients who died before the postoperative PRO data collection timeframe were excluded.

Below we respond to a question from NQF staff:

NQF Question: Please clarify if a claims outcomes analysis of staged procedures was considered to infer potential response rates if included in the populations. In other words, was testing conducted to assess the presence of poor outcomes in the staged procedure population that could equate to not attaining an SCB?

CORE Response: CMS will continue to investigate staged procedures for this measure and work with clinical experts to include as many patients as feasible in the measure cohort through ongoing measure reevaluation and testing. At this time, we determined it was more valuable to move this PRO-PM forward without staged procedures and to incorporate some or all staged procedures into the cohort at a later date as feasible.

[Response Ends]

2b.18. Provide your interpretation of the results, in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results.

In other words, 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.

[Response Begins]

For each exclusion, overall a small percentage of patients (admissions) are removed from the measure, and the exclusions are justified in order to create a fair and balanced measure.

Specifically, the staged procedure exclusion removes a potential negative impact on clinician- and clinician group-specific measure results since the recovery from one procedure may negatively impact recovery from the other procedure. 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. This exclusion represents a small number of the total patients undergoing THA and TKA procedures. This exclusion will be monitored and alternative approaches for including staged procedures will be explored. CMS will continue to consider how to capture staged procedures and appropriately attribute outcomes to the correct procedure, potentially through laterality. However, the current data, both claims and PRO data, are insufficient to provide accurate assessment and attribution.

For patients who died before the postoperative PRO data collection timeframe, it is justified to remove them from the cohort since they were not alive during the postoperative data collection window and unable to provide a response to PROs. Finally, for the exclusion for patients who leave against medical advice, similar to other quality outcome measures, we remove these patients since providers were unable to deliver full care.

[Response Ends]

2b.19. Check all methods used to address risk factors.

[Response Begins]

Statistical risk model with risk factors (specify number of risk factors)

19

[Response Ends]

2b.20. If using statistical risk models, provide detailed risk model specifications, including the risk model method, risk factors, risk factor data sources, coefficients, equations, codes with descriptors, and definitions.

[Response Begins]

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. SCB 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 applied the risk model developed by the hospital-level THA/TKA PRO-PM which was developed 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 the hospital-level THA/TKA PRO-PM measure development team’s technical expert panel (TEP) and clinical experts as both clinically important and feasible.

The risk variables included in the final model are:

- Age, in years
- Male sex
- Body Mass Index (BMI), in kg per m^2
- Procedure: THA
- Bilateral procedure
- Baseline PROMIS Global Mental Health Subscale Score
- 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)
- Pain in Non-Operative Lower Extremity Joint (Total painful joint count: Patient-Reported in Non-operative Lower Extremity Joint) (Ayers et al, 2013)
- Back Pain at preoperative assessment (Quantified Spinal Pain: Patient-Reported Back Pain, Oswestry Disability Index question) (Fairbank et al, 2000; Ayers et al, 2013)
- Narcotic use for >90 days
- 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 clinician- and clinician group-specific RSIR using a hierarchical logistic regression model to account for the natural clustering of observations within clinicians or clinician groups. The model employs a logit link function to link the risk factors to the outcome with a clinician- or clinician group-specific random effect. The risk variable coefficients can be found in the data dictionary (Tab Candidate Risk Variables Included in Risk Modeling).

Let denote the outcome (equal to one if patient has an improvement, zero otherwise) for patient *I* attributed to a clinician or clinician group *j*; denotes a set of risk factors for patient attributed to clinician or clinician group ; and is the number of index admissions attributed to the clinician or clinician group .We assume the outcome is related linearly to the covariates via a logit function:

$$logit(Prob(Y_ij = 1)) = \alpha_j + \beta Z_ij Where\ \alpha_j = \mu + \omega_j; \omega_j\ N(0, \tau^2)$$

where α_j represents the clinician- or clinician group-specific intercept, μ is the adjusted average intercept over all clinicians or clinician groups in the sample, is the clinician- or clinician group-specific intercept deviation from μ, and τ^2 is the between-clinician or clinician group variance component. 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 clinician- and clinician group-specific random effect. The random effects accommodate the assumption that underlying differences in the quality of care across clinicians and clinician groups lead to systematic differences in patient outcomes.

To account for potential response bias, we calculated stabilized inverse probability weights (IPW) 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). We fit the hierarchical logistic regression model to the corresponding parameters along with the stabilized IPW adjust for response bias.

We calculated the clinician and clinician group-specific RSIRs, as the ratio of a clinicians or clinician group's "predicted" number of improvements to "expected" number of improvements multiplied by the overall observed improvement rate. The expected number of improvements for each clinician or clinician group (denominator) was estimated as the sum of the estimated probability of improvement among the clinician's or clinician group's patients accounting for the observed patient characteristics. The predicted number of improvements for each clinician or clinician group (numerator) was estimated as the sum of the estimated probability of improvement of the clinician's or clinician group's patients accounting for the patients' characteristics and the clinician- or clinician group-specific intercept.

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.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.

[Response Ends]

2b.21. If an outcome or resource use measure is not risk-adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (i.e., case mix) is not needed to achieve fair comparisons across measured entities.

[Response Begins]

N/A

[Response Ends]

2b.22. Select all applicable resources and methods used to develop the conceptual model of how social risk impacts this outcome.

[Response Begins]

Published literature

Internal data analysis

[Response Ends]

2b.23. Describe the conceptual and statistical methods and criteria used to test and select patient-level risk factors (e.g., clinical factors, social risk factors) used in the statistical risk model or for stratification by risk.

Please be sure to address the following: potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10 or other statistical tests; correlation of x or higher. Patient factors should be present at the start of care, if applicable. Also discuss any "ordering" of risk factor inclusion; note whether social risk factors are added after all clinical factors. Discuss any considerations regarding data sources (e.g., availability, specificity).

[Response Begins]

In this respecification of the hospital-level THA/TKA PRO-PM (NQF #3559), the risk model developed for the hospital-level THA/TKA PRO-PM team was evaluated for this clinician- and clinician group-level measure. The hospital-level THA/TKA PRO-PM team 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 their Orthopedic Clinical Expert and the TEP and through detailed public comments from orthopedic specialty societies, they 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.

CORE's hospital-level THA/TKA PRO-PM team used the comprehensive list of candidate risk variables obtained through expert and public input to survey their TEP on their thoughts to each risk variable's priority. In addition, they 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 for each clinician and clinician group. 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 for the hospital-level THA/TKA PRO-PM are summarized below:

- The goal of risk adjustment is to account for patient characteristics that are reasonably beyond the control of the clinician. 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 control of the clinician.
    - The goal is not perfect risk prediction – this would imply that the clinician 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. The THA/TKA PRO-PMs can only be implemented by requiring hospitals, surgeons, and patients to collect the PROM and relevant risk variables data both before and after the THA/TKA. The TEP engaged in the development of the hospital-level THA/TKA PRO-PM recommended collection of both a global PROM (the PROMIS Global or VR-12) and a hip- or knee-specific PROM (the HOOS, JR or KOOS, JR). The goal is 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 providers.

Finally, the hospital-level THA/TKA PRO-PM team only included risk variables that were 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 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, the hospital-level THA/TKA PRO-PM team adopted and modified the approach utilized by other quality measures, including the NQF #1550: Hospital-level risk-standardized complication rate (RSCR) following elective primary THA/TKA. The hospital-level THA/TKA PRO-PM team surveyed their TEP and asked them to rank the importance of clinical variables for use in a PRO-PM risk model. They 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. They then assessed model performance in their 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. They 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. They then repeated assessment of model performance for the final combined THA/TKA cohort in the Validation Dataset.

To address non-response bias, the hospital-level THA/TKA PRO-PM team identified variables associated with non-response to PRO survey data in two ways. First, they identified statistical associations of patient characteristics and clinical comorbidities in their data across three PRO response groups: patients with complete PRO data submission, patients with incomplete PRO data submission, and patients with no response. Next, they 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 their data. (See 2b6.1 for a detailed description of the analytic approach to addressing potential response bias).

The conceptual relationship, or potential causal pathways by which social risk factors influence improvement following hip and knee replacement procedures, like the factors themselves, are varied and complex. Similar to other outcome measures, we present four potential pathways that are important to consider:

1. **Patients with social risk factors may have worse health at the time of admission for their surgery**. Patients who have lower income/education or unstable housing may have a worse general health status and may present for their procedure with a greater severity of underlying illness. These social risk factors may contribute to worse health status at admission due to competing priorities, lack of access to care (geographic, cultural, or financial), or lack of health insurance. Given that these risk factors all lead to worse general health status, this causal pathway should be largely accounted for by current clinical risk-adjustment.
2. **Patients with social risk factors often receive care at lower quality hospitals**. Patients of lower income, lower education, or unstable housing have inequitable access to high quality facilities, in part, because such facilities are less likely to be found in geographic areas with large populations of poor patients. Thus, patients with low income are more likely to be seen in lower quality hospitals, which can explain decreased likeihood of achieving the improvement outcome following hospitalization.
3. **Patients with social risk factors may receive differential care within a hospital**. The third major pathway by which social risk factors may contribute to likelihood of not achieving the improvement outcome is that patients may not receive equivalent care within a facility. For example, patients of non-White race or non-English-speaking patients may receive differential or inadequate care and/or education during their stay (such as failure to provide adequate pain control due to biases about pain perception among patients of color or failure to provide educational materials in a patient’s preferred language), leading to poorer health outcomes.
4. **Patients with social risk factors may experience worse health outcomes beyond the control of the health care system.** Some SRFs, such as income or wealth, may affect the likelihood of improvement following hip or knee

replacement without directly affecting health status at admission or the quality of care received during the stay. For instance, while a hospital or provider may make appropriate care decisions and provide tailored care and education, a lower-income patient may have a worse outcome post-discharge due to competing financial priorities which do not allow for adequate recuperation or access to needed treatments, or a lack of access to care outside of the hospital.

Social risk factors often act on multiple pathways, and as such, individual pathways can be complex to distinguish analytically. Some social risk factors, despite having a strong conceptual relationship with worse outcomes, may not have statistically meaningful effects on the risk model. Some social risk factors also have different implications on the decision to risk adjust or not.

Based on this model and because the following factors are currently consistently available in our dataset, the following social risk variables were considered for risk-adjustment:

- Dual-eligible status
  - Following guidance from the Department of Health and Human Services Assistant Secretary for Policy and Evaluation (ASPE) and a body of literature demonstrating differential healthcare and health outcomes among dual eligible patients, we identified dual eligibility as a key variable (ASPE 2016, ASPE 2020). We recognize that Medicare-Medicaid dual eligibility has limitations as a proxy for patient'' income or assets because it does not provide a range of results and is only a dichotomous variable. However, the eligibility threshold for over 65-year-old Medicare patients is valuable, as it takes into account both income and assets and is consistently applied across states for the older population.
- AHRQ SES index
  - The AHRQ SES index score is a well-validated variable that describes the average SES of people living in small defined geographic areas (Bonito et al., 2008). Its value as a proxy for patient-level information is dependent on having the most granular-level data with respect to communities that patients live in. AHRQ-validated SES index score summarizes information from the following 7 variables: percentage of people in the labor force who are unemployed, percentage of people living below poverty level, median household income, median value of owner-occupied dwellings, percentage of people ≥25 years of age with less than a 12th grade education, percentage of people ≥25 years of age completing ≥4 years of college, and percentage of households that average ≥1 people per room.

CMS' decision regarding whether or not to adjust for social risk factors is based both on the empiric results (impact on model and measure scores) and the conceptual model and the use of the measure (in a payment program or for public reporting). In making the decision about whether or not to risk adjust for these factors, CMS also considers the potential unintended consequence of adjusting, and the fairness to patients and providers that care for patients with social risk factors of the unadjusted measure score. If the relationship is driven by poorer quality, adjusting will mask the disparity in care. In contrast, an unadjusted measure will illuminate quality differences and create an incentive to mitigate them. Not adjusting, however may disadvantage providers who care for low SES patients, and unintentionally create an incentive for clinicians to care for fewer patients with social risk factors, potentially reducing access to care. CMS considers this to be a small risk currently, given the correlations between the measure scores calculated with and without social risk factors in the model. CMS also considers alternate approaches to risk adjustment for SRFs, such as stratifying payment based upon performance among peer groups of hospitals caring for similar patients with SRFs. Ongoing research aims to identify valid patient-level social risk factors and highlight disparities related to social risk. As additional variables become available, they will be considered for testing and inclusion within the measure. There are also alternative ways to account for social risk as part of measure program implementation.

For this respecified measure, we evaluated the risk model in our dataset and engaged with stakeholders (clinical expert, Clinical Working Group, Patient Working Group, and TEP). Specifically, 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 risk model in the **Development Dataset**. We then repeated assessment of model performance for the final combined THA/TKA cohort in the **Validation Dataset.**

[Response Ends]

2b.24. Detail the statistical results of the analyses used to test and select risk factors for inclusion in or exclusion from the risk model/stratification.

[Response Begins]

Testing results using the Combined Data of the final risk-adjusted model for SCB improvement following elective primary THA/TKA are presented in Table 15, below. The frequency of risk variables and the risk variable odds ratios are adjusted for other risk variables in the model. 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.

Table 15. Final Risk Model Variables and Adjusted Odds Ratios (Logistic Regression Model): Combined Dataset (Patient N = 19,429)

Risk Factors	Frequency	Odds Ratio (95% CI)
Age Mean (SD)	74 (6.00%)	1 (1.00-1.01)
Male	7,294 (37.54%)	0.82 (0.76-0.87)
BMI Mean (SD)	30 (6.00%)	1.01 (1.00-1.01)
Index admissions with an elective THA procedure	6,971 (35.88%)	1.36 (1.28-1.46)
Number of procedures (two vs. one)	116 (0.60%)	2.07 (1.28-3.34)
Mental Health Score Mean (SD)	50 (8.00%)	0.99 (0.98-0.99)
Health Literacy: Not at all	3,282 (16.89%)	Reference
Health Literacy: A little bit	1502 (7.73%)	1.25 (1.10-1.42)
Health Literacy: Somewhat	2,124 (10.93%)	1.63 (1.45-1.84)

Risk Factors	Frequency	Odds Ratio (95% CI)
Health Literacy: Quite a bit	3,489 (17.96%)	1.74 (1.57-1.93)
Health Literacy: Extremely	9,032 (46.49%)	1.97 (1.81-2.14)
Other Joint Pain: None	6,694 (34.45%)	Reference
Other Joint Pain: Mild	4,768 (24.54%)	0.88 (0.81-0.95)
Other Joint Pain: Moderate	4,897 (25.20%)	0.97 (0.89-1.05)
Other Joint Pain: Severe	2,516 (12.95%)	1.41 (1.26-1.57)
Other Joint Pain: Extreme	554 (2.85%)	2 (1.60-2.50)
Back Pain: None	7,328 (37.72%)	Reference
Back Pain: Very Mild	4,884 (25.14%)	0.95 (0.88-1.03)
Back Pain: Moderate	4,988 (25.67%)	0.93 (0.85-1.00)
Back Pain: Fairly Severe	1,601 (8.24%)	0.95 (0.84-1.07)
Back Pain: Very or Worst Severe	628 (3.23%)	1.48 (1.21-1.81)
Use of Chronic (≥ 90 days) Narcotics	3,390 (17.45%)	0.94 (0.86-1.02)
Severe infection; other infectious diseases (CC 1, 3–7)	3,409 (17.55%)	0.9 (0.83-0.98)
Liver disease (CC 27–31)	813 (4.18%)	0.85 (0.73-0.98)

Risk Factors	Frequency	Odds Ratio (95% CI)
Diabetes mellitus (DM) or DM complications (CC 17-19, 122–123)	5,018 (25.83%)	0.98 (0.91-1.06)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease (CC 40)	2,083 (10.72%)	0.93 (0.84-1.03)
Depression (CC 61)	3,012 (15.50%)	0.92 (0.84-1.01)
Other Psychiatric Disorders (CC 63)	3,099 (15.95%)	0.93 (0.85-1.02)
Coronary atherosclerosis or angina (CC 88–89)	4,750 (24.45%)	0.9 (0.84-0.97)
Vascular or circulatory disease (CC 106–109)	3,727 (19.18%)	0.91 (0.84-0.98)
Renal failure (CC 135–140)	2,753 (14.17%)	1.04 (0.95-1.14)

Below we respond to a question from NQF staff:

NQF Question: Please explain if the health literacy findings are consistent with the general population >= 65 years.

CORE Response: According to the results from the 2003 National Assessment of Adult Literacy, the majority of adults, 53 percent, had *Intermediate* health literacy (Kutner et al., 2006). An additional 22 percent of adults had *Basic* health literacy, 14 percent had *Below Basic* health literacy, and 12 percent had *Proficient* health literacy. Among adults who received Medicare 27% had Below Basic health literacy.

Among procedures with complete PRO and risk variable data (our final cohort), we find 46.5% of patients are in the highest literacy level (extremely comfortable) and 16.9% of patients have the lowest literacy level (not at all comfortable). Compared to the overall Medicare population, the population in our testing dataset had fewer patients with the lowest health literacy levels. We recommend ongoing reevaluation of the measure specifications in broader datasets over time.

Table 2. Health Literacy Responses CJR Dataset (Full Sample)

Health Literacy Response	N (%)
Literacy: Not at all	3282 (16.89%)
Literacy: A little bit	1502 (7.73%)
Literacy: Somewhat	2124 (10.93%)
Literacy: Quite a bit	3489 (17.96%)

Health Literacy Response	N (%)
Literacy: Extremely	1. .49%)

Reference

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy. *National Center for Education Statistics*, *483*, 1–59. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006483

[Response Ends]

2b.25. Describe the analyses and interpretation resulting in the decision to select or not select social risk factors.

Examples may include prevalence of the factor across measured entities, availability of the data source, empirical association with the outcome, contribution of unique variation in the outcome, or assessment of between-unit effects and within-unit effects. Also describe the impact of adjusting for risk (or making no adjustment) on providers at high or low extremes of risk.

[Response Begins]

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 conducted in the **Development Dataset** showed no statistically significant association between AHRQ SES index lowest quartile and SCB improvement, non-White race and SCB improvement, nor dual eligibility and SCB improvement at the bivariate level (Table 16) and when entered into the risk model (Table 17). Tables 18 and 19 provide the mean and range of clinician and clinician group-specific RSIRs with no social risk factors included in the risk model, and with dual eligibility, AHRQ SES index lowest quartile, and non-White race individually included in the risk model. Correlation coefficients for RSIRs calculated without social risk factors with RSIRs calculated individually with each of the social risk factors in the risk model indicate near perfect correlation in our data.

Based on the results of the social risk factor testing, we did not include additional social risk factors beyond health literacy in the risk model. As noted above, we do include health literacy in the final risk model, based upon strong stakeholder support during the development of the hospital-level THA/TKA PRO-PM. In our dataset, neither dual eligibility, AHRQ SES index lowest quartile, nor non-White race were statistically significantly associated with the outcome.

However, similar to the hospital-level THA/TKA PRO-PM 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 16. Bivariate Associations of Social Risk Factors and Race with SCB Improvement: Development Dataset (Patient N = 11,653) (Please note that these categories are not mutually exclusive therefore patients can be counted multiple times in the table.)

Variable	Frequency (%) of Social Risk Factor among Patients in the Development Dataset	Frequency (%) of Social Risk Factor among Patients Achieving SCB Improvement	Frequency (%) of Social Risk Factor among Patients Not Achieving SCB Improvement	P-value
Total	11,653	7,810	3,843	-
Dual eligibility	315 (2.70%)	220 (2.82%)	95 (2.47%)	0.28
AHRQ SES index: **Lowest Quartile**	1,146 (9.83%)	779 (9.97%)	367 (9.55%)	0.48
Race: Non-White	893 (7.66%)	604 (7.73%)	289 (7.52%)	0.68

Cells marked by a dash (-) are intentionally left blank.

Table 17. Adjusted ORs for Social Risk Factors and Race Individually Evaluated in the Risk Model for SCB Improvement: Development Dataset (Patient N = 11,653)

Variable	Frequency (%)	Estimate (Standard Error)	OR (95% CI)	C-Statistic for Model Including Social Risk Factor
Dual eligibility	315 (2.70%)	0.08 (0.13)	1.09 (0.85-1.40)	0.61
AHRQ SES index: **Lowest Quartile**	1,146 (9.83%)	0.04 (0.07)	1.04 (0.91-1.19)	0.61
Race: Non-White	893 (7.66%)	-0.02 (0.08)	0.98 (0.84-1.14)	0.61
Dual eligibility, AHRQ SES Index: **Lowest Quartile, and Race: Non-white included**	-	-	-	0.61

Cells marked by a dash (-) are intentionally left blank.

* C-statistic for the risk model for SCB improvement in the Development Dataset without any of the three social risk factors = 0.609

Table 18. Mean and Distribution of RSIRs Calculated without and with Social Risk Factors and Race in the Risk Model (Development Dataset: Clinicians with >25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	No Additional Social Risk Factors Included	Dual Eligibility	AHRQ SES Index: Lowest Quartile	Race: Non-White	All three social risk factors included
N (Clinicians)	232	232	232	232	232

Summary Statistics	No Additional Social Risk Factors Included	Dual Eligibility	AHRQ SES Index: Lowest Quartile	Race: Non-White	All three social risk factors included
Mean (SD)	64.09% (13.18)	64.10% (13.18)	64.09% (13.18)	64.06% (13.17)	64.08% (13.17)
Percentile	-	-	-	-	-
100% Max	88.41%	88.42%	88.34%	88.20%	88.21%
99%	85.80%	85.83%	85.82%	85.71%	85.78%
95%	82.37%	82.40%	82.29%	81.96%	81.97%
90%	79.53%	79.52%	79.49%	79.67%	79.59%
75% (Q3)	73.44%	73.47%	73.51%	73.27%	73.48%
50% (Median)	66.10%	66.10%	65.99%	66.06%	65.99%
25% (Q1)	55.95%	55.94%	55.97%	55.97%	55.97%
10%	47.77%	47.78%	47.78%	47.59%	47.63%
5%	40.98%	40.99%	40.99%	40.81%	40.95%
1%	22.33%	22.34%	22.31%	22.30%	22.31%
0% Min	18.45%	18.45%	18.44%	18.37%	18.38%
Pearson Correlation Coefficient (With “No Social Risk Factors”)	-	>0.99	>0.99	>0.99	>0.99

Cells marked by a dash (-) are intentionally left blank.

Table 19. Mean and Distribution of RSIRs Calculated without and with Social Risk Factors and Race in the Risk Model (Development Dataset: Clinician Groups with ≥25 THA/TKA Patients with Complete PRO Data)

Summary Statistics	No Risk Factors Included	Dual Eligibility	AHRQ SES Index: Lowest Quartile	Race: Non-White	All three social risk factors included
N (Clinician Groups)	170	170	170	170	170
Mean (SD)	64.59% (12.77)	64.59% (12.77)	64.56% (12.78)	64.49% (12.75)	64.50% (12.75)
Percentile	-	-	-	-	-
100% Max	86.08%	86.09%	86.19%	86.25%	86.35%
99%	85.34%	85.35%	85.25%	85.04%	85.06%
95%	81.30%	81.29%	81.31%	81.36%	81.29%
90%	79.74%	79.74%	79.65%	79.38%	79.36%
75% (Q3)	73.24%	73.25%	73.28%	73.07%	73.14%
50% (Median)	66.57%	66.56%	66.47%	66.27%	66.21%
25% (Q1)	57.43%	57.42%	57.45%	57.90%	57.87%
10%	46.67%	46.68%	46.61%	46.73%	46.65%
5%	39.06%	39.06%	39.13%	38.97%	39.08%
1%	21.59%	21.60%	21.60%	21.51%	21.58%
0% Min	21.42%	21.42%	21.56%	21.37%	21.50%
Pearson Correlation Coefficient (With "No Social Risk Factors")	-	>0.99	>0.99	>0.99	>0.99

Cells marked by a dash (-) are intentionally left blank.

Below we respond to questions from NQF staff:

NQF Question 1: In Table 15, the tested sample is >92% White. Please address whether enough non-white patients were sampled to determine if race should be included in the risk model.

CORE Response: We examined elective, primary THA/TKA patients in the Medicare FFS population between April 2017-March 2020 and found 9% of patients in the cohort were non-White and 4% of patients were dually eligible for Medicare and Medicaid; these results are consistent with known disparities in offer and acceptance rates for THA/TKA among non-White patients. In our measure testing dataset, we found slightly lower percentages of non-White (7.6%) and dually eligible patients (2.7%) than seen nationally. Since PROMs are not systematically captured on our target population at the national level, we utilized the CJR dataset for testing and recommend ongoing evaluation of the risk model in the future. Given the known variation in response rates to PROs due to social risk factors, our statistical approach to potential response bias applies weighting based on important factors such as race and dual eligibility (as well as Agency for Healthcare Research and Quality [AHRQ] socioeconomic [SES] index).

In addition, the risk variable included in the model were selected a priori from the orthopedic community and extensive vetting.

NQF Question 2: The developer states, "Due to known associations between race and poorer outcomes, we also assessed the association between non-White race and the outcome." As the tested sample is >92% White, please explain if enough non-White patients were sampled to determine if race should be included in the risk model.

CORE Response: Please see the answer above.

[Response Ends]

2b.26. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (describe the steps—do not just name a method; what statistical analysis was used). Provide the statistical results from testing the approach to control for differences in patient characteristics (i.e., case mix) below. If stratified ONLY, enter “N/A” for questions about the statistical risk model discrimination and calibration statistics.

Validation testing should be conducted in a data set that is separate from the one used to develop the model.

[Response Begins]

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)

(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

[Response Ends]

2b.27. Provide risk model discrimination statistics.

For example, provide c-statistics or R-squared values.

[Response Begins]

Model performance statistics for the risk model for meeting or exceeding the SCB improvement threshold are provided in Table 20.

In the Development Dataset:

- C-statistic for the risk model is 0.61
- predictive ability from the lowest to highest decile is 48.67%- 80.03%

In the Validation Dataset:

- C-statistic for the risk model is 0.607

predictive ability from the lowest to highest decile is 52.44%- 81.14%

Table 20. Model Performance of Risk-Adjusted Model of SCB Improvement following THA/TKA

Model Performance Statistic	Development Dataset	Validation Dataset
C-statistic	0.61	0.607
Discrimination- Predictive ability (lowest decile %- highest decile %)	(48.67%, 80.03%)	(52.44%, 81.14%)

[Response Ends]

2b.28. Provide the statistical risk model calibration statistics (e.g., Hosmer-Lemeshow statistic).

[Response Begins]

The calibration indices (γ0, γ1) used to assess the risk model for meeting or exceeding SCB improvement are for the Validation Dataset are (0.0216, 0.9733).

[Response Ends]

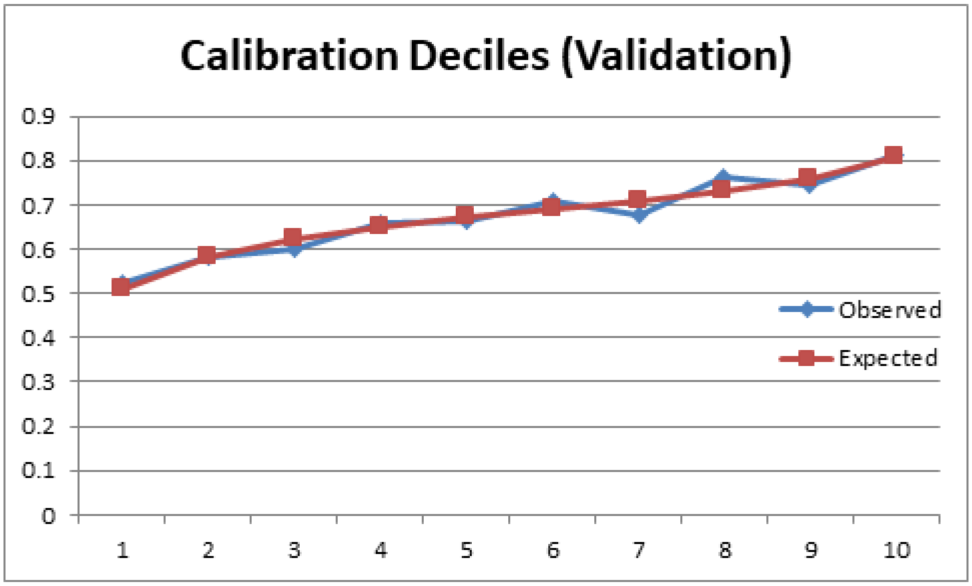
2b.29. Provide the risk decile plots or calibration curves used in calibrating the statistical risk model.

The preferred file format is .png, but most image formats are acceptable.

[Response Begins]

Figure 5 plots risk deciles for the **Validation Dataset**.

Figure 5. Calibration Deciles for the Validation Dataset



[Response Ends]

2b.30. Provide the results of the risk stratification analysis.

[Response Begins]

N/A

[Response Ends]

2b.31. Provide your interpretation of the results, in terms of demonstrating adequacy of controlling for differences in patient characteristics (i.e., case mix).

In other words, what do the results mean and what are the norms for the test conducted?

[Response Begins]

The following results demonstrate that the risk-adjustment model adequately controls for differences in patient characteristics:

Results demonstrate the risk-adjustment model moderately controls for differences in patient characteristics.

Discrimination statistics

The calculated C-statistic was 0.61 using the **Development Dataset** and 0.607 using the **Validation Dataset** and indicates adequate model discrimination across the cohort models. With both the **Development** and **Validation** Datasets, the model indicated a moderate range between the lowest decile and highest decile, indicating the ability to distinguish high-risk subjects from low-risk subjects.

-

Calibration statistics (γ0, γ1)

The calibration values which are consistently close to zero 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 in the **Validation Dataset**.

Risk Decile Plot

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

Overall Interpretation

Interpreted together, our diagnostic results demonstrate that the risk-adjustment model moderately controls for differences in patient characteristics (case mix) and bias due to non-response.

[Response Ends]

2b.32. Describe any additional testing conducted to justify the risk adjustment approach used in specifying the measure.

Not required but would provide additional support of adequacy of the risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed.

[Response Begins]

N/A

[Response Ends]

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.

3.01. Check all methods below that are used to generate the data elements needed to compute the measure score.

[Response Begins]

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

[Response Ends]

3.02. Detail to what extent the specified data elements are available electronically in defined fields.

In other words, indicate whether data elements that are needed to compute the performance measure score are in defined, computer-readable fields.

[Response Begins]

Patient/family reported information (may be electronic or paper)

[Response Ends]

3.03. 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 data elements not from electronic sources.

[Response Begins]

Currently, this measure allows clinicians and clinician groups to collect data using a range of methods, including paper and electronic formats. The measure uses patient-reported data for the outcome definition and patient- and provider-reported data and administrative claims data for the risk model. The PRO and clinical risk variable data were not electronically specified in the measure development and testing datasets; most, if not all, clinical data elements can feasibly be captured in the electronic health record as they represent standardized results that can be captured in discrete fields. Leveraging administrative claims data to augment limited clinical risk variables allows the measure to capture prior medical history and comorbidities without increased patient or provider burden.

While we strongly support the use of electronic data capture, not all clinicians collect patient-reported outcomes on patients undergoing elective primary THA/TKA procedures and fewer collect these data in electronic form. 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 electronic clinical quality measure (eCQM) process measures (Functional Status Outcomes for Patients Receiving Primary Total Hip Replacements and Functional Status Outcomes for Patients Receiving Primary Total Knee Replacements) that incentivize collection of the PRO data needed to calculate the measure outcome, making future e-specification less burdensome.

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

N/A

[Response Ends]

3.06. 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.

[Response Begins]

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 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. Patients have expressed to us the importance of knowing what PRO survey results will be used for and noted a greater willingness to complete surveys if they are collected by their provider. In regard to data collection barriers, we heard interest from providers to have sufficient time and resources for the initial set up of PRO implementation infrastructure and processes. They noted that PROM capture either remotely or in-person is resource intensive and the cost of hiring external vendors to support PROM data capture is high.

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.

[Response Ends]

Consider implications for both individuals providing data (patients, service recipients, respondents) and those whose performance is being measured.

3.07. Detail 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),

Attach the fee schedule here, if applicable.

[Response Begins]

N/A

[Response Ends]

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.

Extent to which intended audiences (e.g., consumers, purchasers, providers, policy makers) can understand the results of the measure and are likely to find them useful for decision making.

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 demonstrating performance improvement.

4a.01.

Check all current uses. For each current use checked, please provide:

Name of program and sponsor

URL

Purpose

Geographic area and number and percentage of accountable entities and patients included

Level of measurement and setting

[Response Begins]

Not in use

This PRO-PM is being submitted for initial endorsement and is not currently used in any accountability program. While CMS has not formally proposed the measure for a specific program, we understand CMS' intent is to publicly report the measure results.

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]

Public reporting

Payment Program

[Response Ends]

4a.03. If not currently publicly reported OR used in at least one other accountability application (e.g., payment program, certification, licensing), explain why the measure is not in use.

For example, do policies or actions of the developer/steward or accountable entities restrict access to performance results or block implementation?

[Response Begins]

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

[Response Ends]

4a.04. If not currently publicly reported OR used in at least one other accountability application, provide a credible plan for implementation within the expected timeframes: used in any accountability application within 3 years, and publicly reported within 6 years of initial endorsement.

A 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.

[Response Begins]

CMS may opt to implement this measure in the Quality Payment Program (QPP) through rulemaking in the future.

[Response Ends]

4a.05. Describe how performance results, data, and assistance with interpretation have been provided to those being measured or other users during development or implementation.

Detail 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.

[Response Begins]

This PRO-PM has not been implemented yet and thus measure results have not been shared with the measured entities (clinicians and clinician groups). However, feedback was obtained from a TEP (21 total members, five of which were patients), a Clinical Working Group (four clinical expert members representing each of the four national THA and/or TKA professional societies), and a Patient Working Group (six members). TEP members were selected through a publicly posted call for TEP on the CMS website and patients were recruited through partnerships with Rainmakers (CMS' contracted person and family engagement contractor). Clinical Working Group members were nominated by the American Academy of Orthopaedic Surgeons (AAOS), American Association of Hip and Knee Surgeons (AAHKS), the Hip Society, and the Knee Society. Feedback was obtained via teleconference calls. 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.02 (Provide evidence that the target population values the measured outcome, process, or structure and finds it meaningful) of this form.

[Response Ends]

4a.06. Describe the process for providing measure results, including when/how often results were provided, what data were provided, what educational/explanatory efforts were made, etc.

[Response Begins]

Throughout measure respecification, we have engaged the TEP, Clinical Working Group, and Patient Working Group. To date, the TEP has provided input on and supported the measure concept, clinician, and clinician group attribution of THA/TKA procedures, and risk model approach and results. In addition, we reviewed the approach to social risk factor analyses and results, approach to response bias and results, the final measure scores and reliability and validity testing. We also reviewed future measure specification updates, such as expanding the measure cohort and extending the postoperative PROM data collection window. The Clinical Working Group was consulted on and supported the measure concept, the risk model, and risk model results. We also reviewed the final measure scores and reliability and validity testing. In addition, we asked the Clinical Working Group about the final measure results and analyses related to future measure specification updates. The Patient Working Group provided input on and supported the measure concept, measure use, and approach to analyzing social risk and non-response bias. We also discussed future measure specification updates. Statistical analyses were shared with the TEP, Clinical Working Group, and Patient Working Group. We assessed face validity by asking the TEP and Patient Working Group members to rate the measure according to two statements.

[Response Ends]

4a.07. Summarize the feedback on measure performance and implementation from the measured entities and others. Describe how feedback was obtained.

[Response Begins]

Feedback was obtained via four teleconference meetings with the TEP, three teleconference meetings with the Clinical Working Group, and three teleconference meetings with the Patient Working Group. The TEP and Clinical Working Group indicated strong support of measure specifications and provided recommendations for ongoing evaluation, such as consideration of provider volume, handling of staged procedures, the impact of social risk, and the expansion of the postoperative timeframe.

[Response Ends]

4a.08. Summarize the feedback obtained from those being measured.

[Response Begins]

The TEP, which includes multiple clinicians, the Clinical Working Group, comprised entirely of four clinicians, and our clinical expert indicated strong support for a clinician- and/or clinician group-level measure of patient-reported outcomes following elective primary THA/TKA. They recommended ongoing evaluation of the risk model and social risk factor analyses.

[Response Ends]

4a.09. Summarize the feedback obtained from other users.

[Response Begins]

The Patient Working Group members indicated strong support for a patient-reported, outcomes-based performance measure following elective THA and TKA. They indicated that a clinician-level measure would be most useful in selecting their surgeon, and that a clinician group-level measure would also be helpful in making informed decisions. The Patient Working Group also expressed interest in gaining access to additional outcome rates alongside this measure, such as complication and infection rates. They supported consideration of health equity variables in future evaluations of the risk model. Additionally, the Patient Working Group recognized that although individuals with high PROM scores before their THA/TKA procedure (indicating less severity preoperatively) are less likely to reach the substantial clinical benefit (SCB) improvement threshold, and they are important to include in this measure.

[Response Ends]

4a.10. Describe how the feedback described has been considered when developing or revising the measure specifications or implementation, including whether the measure was modified and why or why not.

[Response Begins]

TEP, Clinical Working Group, and Patient Working Group feedback has been considered throughout measure respecification. Furthermore, the hospital-level THA/TKA PRO-PM development team engaged with patients during the selection of the cohort, measure outcome, data collection instruments, and risk adjustment model.

[Response Ends]

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, 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, provide an explanation. 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.

[Response Begins]

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 clinicians and/or clinician groups with performance information necessary to implement focused quality improvement efforts.

[Response Ends]

4b.02. Explain any unexpected findings (positive or negative) during implementation of this measure, including unintended impacts on patients.

[Response Begins]

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

[Response Ends]

4b.03. Explain any unexpected benefits realized from implementation of this measure.

[Response Begins]

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

[Response Ends]

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.

If you are updating a maintenance measure submission for the first time in MIMS, please note that the previous related and competing data appearing in question 5.03 may need to be entered in to 5.01 and 5.02 if the measures are NQF endorsed. Please review and update questions 5.01, 5.02, and 5.03 accordingly.

5.01. Search and select all NQF-endorsed related measures (conceptually, either same measure focus or target population).

(Can search and select measures.)

[Response Begins]

0422: Functional status change for patients with Knee impairments

0425: Functional Status Change for Patients with Low Back 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)

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

0423: Functional status change for patients with Hip impairments

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

0426: Functional status change for patients with Shoulder impairments

0428: Functional status change for patients with General orthopaedic impairments

3461: Functional Status Change for Patients with Neck Impairments

3559: Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA)

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

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

[Response Ends]

5.02. Search and select all NQF-endorsed competing measures (conceptually, the measures have both the same measure focus or target population).

(Can search and select measures.)

[Response Begins]

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

[Response Ends]

5.03. If there are related or competing measures to this measure, but they are not NQF-endorsed, please indicate the measure title and steward.

[Response Begins]

The measure aligns with the electronic clinical quality process measures which incentivize the collection 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) for elective primary THA and TKA procedures, respectively. The measure names are Functional Status Assessment for Total Hip Replacement (QPP Quality ID: 376) and Functional Status Assessment for Total Knee Replacement (QPP Quality ID: 375). The measure steward for these two measures is CMS.

[Response Ends]

5.04. If this measure conceptually addresses EITHER the same measure focus OR the same target population as NQF-endorsed measure(s), indicate whether the measure specifications are harmonized to the extent possible.

[Response Begins]

Yes

[Response Ends]

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

[Response Begins]

This PRO-PM differs from NQF #2653 in cohort, outcome, and risk adjustment.

Cohort: This PRO-PM includes both THA and TKA procedures, as clinical experts agree that clinician-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 procedure. 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. The timing of PRO data collection is 90 – 0 days prior to and 270 – 365 days following the procedure. 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 clinicians and clinician groups the risk-adjusted proportion of patients with improvement. In contrast, NQF #2653 collects PRO data with the Oxford Knee Score three months prior to and 9 – 15 months following the procedure 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 the hospital-level THA/TKA PRO-PM development 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. Furthermore, the TEP and Patient Working Group convened during development of the hospital-level THA/TKA PRO-PM stated concerns with measuring an average change score because entities with all average outcomes would look similar to entities whose patients either did very well or very poorly (bimodal distributed outcomes), thus providing potentially misleading information to consumers and patients.

Risk Adjustment: The risk model for this PRO-PM includes important risk variables, supported by the hospital-level THA/TKA PRO-PM development 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 the American Association of Hip and Knee Surgeons (AAHKS) and the American Academy of Orthopedic Surgeons (AAOS) through public comment (Centers for Medicare & Medicaid Services, CJR Final Rule 2015, Section III.D.3.A).

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).

[Response Ends]

5.06. Describe why this measure is superior to competing measures (e.g., a more valid or efficient way to measure quality). Alternatively, justify endorsing an additional measure.

Provide analyses when possible.

[Response Begins]

This PRO-PM is superior to NQF #2653 for the following reasons: 1) it assesses SCB improvement with a binary outcome that elucidates for clinicians, clinician groups, and patients the risk-adjusted proportion of patients with improvement (a clear, understandable metric that patients support); 2) it uses a more robust and stakeholder-driven risk model, anticipated to produce a measure with greater face validity with stakeholders; and 3) it is harmonized with related measures including NQF #3559 Hospital-Level, Risk-Standardized Patient-Reported Outcomes Following Elective Primary Total Hip and/or Knee Arthroplasty (THA/TKA), NQF #3493 Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) for Merit-based Incentive Payment System (MIPS) Eligible Clinicians and Eligible Clinician Groups, and NQF #1550 Hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and total knee arthroplasty (TKA).

[Response Ends]

Appendix

Supplemental materials may be provided in an appendix.: Available in attached file

Attachment: 3639_QPPMeasureMethodologyReport_ForPublicComment_09.17.21_FINAL.pdf

Contact Information

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

Measure Steward Point of Contact: Poyer, James, james.poyer@cms.hhs.gov

Measure Developer if different from Measure Steward: Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE)

Measure Developer Point(s) of Contact: Vellanky, Smitha, smitha.vellanky@yale.edu

Sutton, Lamont, doris.peter@yale.edu

Additional Information

1. Provide any supplemental materials, if needed, as an appendix. All supplemental materials (such as data collection instrument or methodology reports) should be collated one file with a table of contents or bookmarks. If material pertains to a specific criterion, that should be indicated.

[Response Begins]

Available in attached file

[Response Ends]

Attachment: 3639_QPPMeasureMethodologyReport_ForPublicComment_09.17.21_FINAL.pdf

2. List the workgroup/panel members' names and organizations.

Describe the members' role in measure development.

[Response Begins]

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

1. Lisa G. Suter, MD – Contract Director and Project Director. Provided experience relevant to clinical content and performance measurement.

2. Kathleen M.B. Balestracci, PhD – New Measure Division Lead. Provided experience relevant to performance measurement.

3. Kerry McDowell, MS, MPhil – Project Manager. Provided experience relevant to performance measurement.

5. Zhenqiu Lin, PhD – Analytic Director. Provided experience relevant to performance measurement.

6. Sheng Zhou, MD, ScM – Lead Analyst. Provided experience relevant to performance measurement.

7. Kyaw (Joe) Sint, PhD, MPH – Supporting Analyst. Provided experience relevant to performance measurement.

8. Rachelle Zribi, BA –Project Lead. Provided experience relevant to performance measurement.

9. Jasie Mathew, MBA- Project Coordinator. Provided experience relevant to performance measurement.

9. Shani Legore, BA – Person and Family Engagement Communication Specialist. Provided experience relevant to Patient Working Group facilitation and performance measurement.

10. Emma Turchick, MPH – Research Associate. Provided experience relevant to performance measurement.

11. Matthew Saenz – Consultant. Provided experience relevant to performance measurement.

Technical Expert Panel (TEP) Members

1. David C. Ayers, MD – Professor and Chair of Orthopaedics and Physical Rehabilitation, University of Massachusetts (UMass) Medical School. Provided experience relevant to clinical content and performance measurement.

2. Thomas C. Barber, MD – Deputy Physician in Chief, Memorial Sloan Kettering Hospital. Provided experience relevant to clinical content and performance measurement.

3. Phyllis Bass - Patient Expert. Recipient of elective THA or TKA procedure. Provided patient perspective.

4. Vinod Dasa, MD – Professor of Clinical Orthopedics and Director of Research, Louisiana State University Health Sciences Center. Provided experience relevant to clinical content and performance measurement.

5. Rachel DuPré Brodie – Senior Director of Measurement & Accountability, Purchaser Business Group on Health (PBGH). Provided experience relevant to performance measurement.

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

7. William G. Hamilton, MD – Chair of Orthopedic Surgery, Inova Mount Vernon Hospital; FOCAL Chair, American Association of Hip and Knee Surgeons. Provided experience relevant to clinical content and performance measurement.

8. Cynthia S. Jacelon, PhD, RN-BC, CRRN, FGSA, FAAN – Professor and Executive Associate Dean, University of Massachusetts Amherst School of Nursing. Provided experience relevant to clinical content and performance measurement.

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

10. Patient B – Patient Expert. Recipient of elective THA or TKA procedure. Provided patient perspective.

11. Craig T. Miller, PT – Director of Home Care Therapy and Senior PT, Rivetus Rehabilitation and American Physical Therapy Association. Provided experience relevant to clinical content and performance measurement.

12. Michael H. Perskin, MD – The American Geriatrics Society; Clinical Professor of Medicine, New York University School of Medicine. Provided experience relevant to clinical content and performance measurement.

13. Nan Rothrock, PhD – Associate Professor of Medical Social Sciences, Feinberg School of Medicine of Northwestern University. Provided experience relevant to clinical content and performance measurement.

14. Jonathan L. Schaffer, MD, MBA, FACS, FHIMSS, FABOS – Staff and Program Director, The Cleveland Clinic. Provided experience relevant to clinical content and performance measurement.

15. Adam Schwartz, MD, MBA – Associate Professor of Orthopaedic Surgery, Mayo Clinic. Provided experience relevant to clinical content and performance measurement.

16. Robert Sterling, MD – Associate Professor of Orthopaedic Surgery and Vice Chair for Quality, Safety, and Service, Johns Hopkins University School of Medicine. Provided experience relevant to clinical content and performance measurement.

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

18. Christine Von Raesfeld – Patient Expert. Recipient of elective THA or TKA procedure. Provided patient perspective.

19. Patricia Walker, PhD – Patient Expert. Recipient of elective THA or TKA procedure. Provided patient perspective.

20. Kevin Woodward, PA-C, MMS – Physician Assistant of Orthopaedic Surgery, American Academy of Physician Assistants, Maryland Academy of Physician Assistants, John Hopkins University. Provided experience relevant to clinical content and performance measurement.

21. Adolph J. Yates, Jr, MD, FAAOS, FAOA – Chief of Orthopedic Surgery, UPMC-Shadyside Hospital; Professor and Vice Chair for Quality, Department of Orthopedic Surgery, University of Pittsburgh School of Medicine. Provided experience relevant to clinical content and performance measurement.

Clinical Working Group

1. James I. Huddleston, III, MD – Associate Professor, Department of Orthopaedic Surgery, Stanford University Medical Center; Chief of Arthritis Service, Department of Orthopaedic Surgery, Stanford University Medical Center. Provided experience relevant to clinical content.

2. Jay R. Lieberman, MD – Professor and Chair, Department of Orthopaedic Surgery, Director of Institute of Orthopaedics, Keck School of Medicine of the University of Southern California; Professor of Biomedical Engineering, Viterbi School of Engineering of the University of Southern California; Second Vice President, The Hip Society. Provided experience relevant to clinical content.

3. Mary I. O'Connor, MD – Chief Medical Office, Vori Health. Provided experience relevant to clinical content.

4. Kathryn Schabel, MD – Associate Professor of Orthopaedic Surgery, Adult Reconstruction, Oregon Health and Science University. Provided experience relevant to clinical content.

Patient Working Group

1. Earl Shellner – Recipient of elective THA or TKA procedure. Provided patient perspective.

2. Linda Radach – Recipient of elective THA or TKA procedure. Provided patient perspective.

3. Rosie Bartel – Recipient of elective THA or TKA procedure. Provided patient perspective.

4. Richard Duncan – Recipient of elective THA or TKA procedure. Provided patient perspective.

5. Barbra Kivowitz – Recipient of elective THA or TKA procedure. Provided patient perspective.

6. Suzanne Nevins – Recipient of elective THA and TKA procedure. Provided patient perspective.

CORE Expert Clinical Consultant

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

[Response Ends]

3. Indicate the year the measure was first released.

[Response Begins]

N/A

[Response Ends]

4. Indicate the month and year of the most recent revision.

[Response Begins]

N/A

[Response Ends]

5. Indicate the frequency of review, or an update schedule, for this measure.

[Response Begins]

N/A

[Response Ends]

6. Indicate the next scheduled update or review of this measure.

[Response Begins]

N/A

[Response Ends]

7. Provide a copyright statement, if applicable. Otherwise, indicate “N/A”.

[Response Begins]

N/A

[Response Ends]

8. State any disclaimers, if applicable. Otherwise, indicate “N/A”.

[Response Begins]

N/A

[Response Ends]

9. Provide any additional information or comments, if applicable. Otherwise, indicate “N/A”.

[Response Begins]

N/A

[Response Ends]