Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measures (PRO-PMs)

Version 1.0 DRAFT Methodology Report

Submitted by:
Yale New Haven Health Services Corporation — Center for Outcomes Research and Evaluation (YNHHSC/CORE)

Prepared for:
Centers for Medicare & Medicaid Services (CMS)

October 2021

Table of Contents

List of Tables

List of Figures

Yale New Haven Health Services Corporation — Center for Outcomes Research and Evaluation (CORE) Project Team

Rachelle Zribi, BA – Project Lead
Kathleen M.B. Balestracci, PhD, MSW* – New Measure Division Lead
Zhenqiu Lin, PhD – Analytic Director
Sheng Zhou, ScM, MD* – Lead Analyst
Kyaw (Joe) Sint, PhD, MPH* – Supporting Analyst
Emma Turchick, MPH – Research Associate I
Kerry McDowell, MS, MPhil – Project Manager
Lisa G. Suter, MD* – Contract Director and Project Director
*Yale School of Medicine

Acknowledgements

This work is a collaborative effort, and the authors gratefully acknowledge the individuals that provided feedback, experiences, and advice on the development of this measure: former team members Mariana Henry, MPH, Fior Rodriguez, BS, and Miriam Katz, MPH, consultants X4 and Matthew Saenz, our Technical Expert Panel (TEP) members, our Clinical Working Group, our Patient Working Group, our Clinical Expert, Kevin Bozic, MD, MBA, the Person and Family Engagement team, the hospital-level THA/TKA PRO-PM development team, and the clinician-level THA/TKA Complication development team. Acknowledgment of the members of our TEP, Clinical Working Group, and Patient Working Group are offered in Appendix A. We would like to thank our Contracting Officer Representative at the Centers for Medicare & Medicaid Services (CMS), James Poyer, MS, MBA for his continued support and contributions to our work.

Disclaimer: The views, thoughts, and opinions expressed in this report belong solely to the authors, and not necessarily to any contributors or consultants, including the Patient Working Group, Clinical Working Group, and Technical Expert Panel members and their affiliated organizations. Acknowledgement of input does not imply endorsement of the methodology and policy decisions.

Executive Summary

Measure Background

Elective total hip arthroplasty and total knee arthroplasty (THA and TKA, respectively) are important, effective procedures performed on a broad population. They offer significant improvement in quality of life by reducing pain and improving function and mobility for the majority of patients undergoing these procedures. They are costly and frequently performed surgeries, most commonly performed for degenerative joint disease and osteoarthritis, conditions affecting millions of Americans. As such, they are priority areas for patient-reported outcome performance measure (PRO-PM) development.

The Centers for Medicare & Medicaid Services (CMS) contracted with Yale New Haven Health Services Corporation — Center for Outcomes Research and Evaluation (CORE) to develop an eligible clinician-level and/or eligible clinician group-level PRO-PM that reflects the quality of care for patients undergoing elective primary THA and TKA. CORE is developing the measure for use under the Quality Payment Program (QPP) for eligible clinicians in the Merit-based Incentive Payment System (MIPS). This measure is a re-specification of the existing, hospital-level THA/TKA PRO-PM[1] (National Quality Forum-endorsed measure #3559) and has been designed to align with the measure specifications of that measure. This report presents the approach to development, specifications, and testing results of the clinician-level and/or clinician group-level THA/TKA PRO-PM (hereafter MIPS THA/TKA PRO-PM). Measure development has benefited from close stakeholder engagement, including a nationally convened TEP, a Clinical Working Group, and a Patient Working Group.

Current clinician-level quality improvement measures for patients undergoing elective THA and TKA procedures are focused primarily on complications and evidence-based processes of care. Measurement of patients' self-assessment of their health, such as pain and function, allows for a direct way to capture patients' experience of care and their results. Patient-reported outcomes (PROs) assessing health status as a result of care are a critical type of outcome needed for healthcare quality assessment.

Measure Development

In this report, we present the development of the MIPS THA/TKA PRO-PM and measure specifications. We used a multi-faceted approach to develop measure specifications, including consultations with experts experienced in the collection and use of THA/TKA PRO data, prospective data collection and analyses to inform measure development and feasibility, and stakeholder engagement in the form of a national TEP, a Clinical Working Group, an orthopedic expert, and a Patient Working Group.

Due to the absence of large scale, and uniformly collected available PRO data from patients undergoing elective primary THA/TKA and in alignment with the hospital-level THA/TKA PRO-PM, we use data incentivized and voluntarily collected within the Center for Medicare and Medicaid Innovation's (CMMI's) Comprehensive Care for Joint Replacement (CJR) model to support measure development and testing. Details on the CJR Model can be found at https://innovation.cms.gov/innovation-models/cjr.

Measure specifications have been finalized, the risk model has been developed, and testing has been completed. Assessments of social risk factors for inclusion in the risk model and analyses for addressing potential bias due to non-response were conducted. A more detailed description and rationale for measure decisions are provided in the body of the report.

Measure Specifications

Data Sources: The MIPS THA/TKA PRO-PM primarily uses PRO data collected from patients undergoing eligible THA/TKA procedures, risk variable data collected from these patients and their providers, and administrative claims data for Medicare fee-for-service (FFS) beneficiaries. PRO and additional risk variable data are collected preoperatively, and PRO data are collected again postoperatively on patients undergoing an elective primary THA/TKA. Claims data are used to identify eligible elective primary THA/TKA procedures for the measure cohort and candidate risk variables, including patient demographics and clinical comorbidities up to 12 months prior to the procedure. Medicare Part A inpatient data and Medicare Part B inpatient claim and claim line data from the Integrated Data Repository (IDR) are used to match patients to clinicians and clinician groups who billed for the procedure. Three additional data sources provide data for the measure: the Medicare Enrollment Database (EDB) identifies Medicare FFS enrollment and race, the Master Beneficiary Summary File (MBSF) allows for determination of dual eligibility status, and the American Community Survey data allow for derivation of the Agency for Healthcare Research and Quality (AHRQ) socioeconomic status (SES) index score. Data from these sources are linked for patients undergoing elective primary THA and TKA procedures for the measurement period.

Measure Cohort: The measure cohort includes Medicare FFS patients 65 years of age or older undergoing elective primary THA/TKA procedures. Patients with fractures and revisions are excluded from the measure cohort. The measure cohort is intentionally aligned with CMS's existing hospital-level THA/TKA PRO-PM and 90-day complication measure cohorts and CMS's Clinician and Clinician Group Risk-standardized Complication Rate Following Elective Primary Total Hip Arthroplasty and/or Total Knee Arthroplasty measure cohort (hereafter referred to as the MIPS THA/TKA Complication measure).

Clinician Attribution: We used the approach developed for the MIPS THA/TKA Complication measure to match patients to eligible clinicians and/or clinician groups. We made one enhancement to limit the eligible clinician specialty types to further align with providers primarily responsible for the THA/TKA procedure and surgical outcomes.

Measure Outcome: The numerator is the risk-standardized proportion of patients undergoing an elective primary THA or TKA who meet or exceed a patient-defined substantial clinical benefit (SCB) threshold of improvement from preoperative to postoperative assessments on joint-specific patient-reported outcome measures (PROMs).

The measure outcome will assess patient improvement in PROs following elective primary THA/TKA. Patient improvement will be measured using the joint-specific instruments chosen for CJR PRO data

collection and vetted through the hospital-level THA/TKA PRO-PM stakeholder engagement during development:

- The Hip dysfunction and Osteoarthritis Outcome Score for Joint Replacement (HOOS, JR)[2] for THA patients and
- The Knee injury and Osteoarthritis Outcome Score for Joint Replacement (KOOS, JR)[3] for TKA patients.

We recommend PRO data be collected 90 to zero days prior to the procedure and 300 to 425 days following the procedure.

- The postoperative data collection period finalized in the CJR model was 270 to 365 days after the procedure, which was used in the development and testing of this measure. Extensive input was provided by clinical experts following measure development, however, strongly recommending a refinement to the postoperative data collection period to better align with clinical workflow and typical one-year follow-up scheduling and to allow for better postoperative PRO data capture. Based on this input, we propose a postoperative PRO data collection period of 300 to 425 days after the procedure. We anticipate, based upon extensive stakeholder input, this will result in limited impact to the measure's scientific acceptability while significantly increasing clinical acceptance and response rates.

The measure outcome defines patient improvement as a binary outcome (yes/no) of meeting or exceeding an SCB between preoperative and postoperative assessments on the joint-specific PROMs as follows:

- For THA patients: meeting or exceeding the SCB threshold of 22 points on the HOOS, JR.
- For TKA patients: meeting or exceeding the SCB threshold of 20 points on the KOOS, JR.

Risk Adjustment: We aligned the risk-adjustment variables with those used with the hospital-level THA/TKA PRO-PM. Details on the consensus-based approach to developing the risk model can be found in Section 2.6 of the Hospital-level THA/TKA PRO-PM Measure Methodology Report Version 1.0.

Measure Calculation: Clinician-and clinician group-level specific risk-standardized improvement rates (RSIRs) are calculated, producing a performance measure per clinician or clinician group which accounts for patient case mix and represents a measure of quality of care following primary elective THA and TKA.

Addressing Potential Response Bias: We aligned the approach to missing PRO data and potential bias in the measure with the hospital-level THA/TKA PRO-PM approach. In summary, the hospital-level THA/TKA PRO-PM development team conducted a thorough literature search and identified several approaches for statistically addressing potential nonresponse bias (e.g., covariates adjustment in regression, submission score adjustment in regression, and stabilized inverse propensity score weighted regression). Following consultation with a statistical expert, the team decided to address 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.

Testing: CORE conducted testing of the MIPS THA/TKA PRO-PM with PRO and risk variable data collected from CJR participant hospitals, matched to CMS administrative claims data, and matched to eligible clinicians and clinician groups. CORE calculated clinician- and clinician group-level RSIRs and found RSIRs ranged from 18.36% to 88.56% (median: 65.75%) for clinicians and 20.86% – 85.90% (median: 66.69%) for clinician groups. CORE conducted signal-to-noise measure score reliability testing and found signal-to-noise ratio sufficiently high (0.87 for clinicians and 0.92 for clinician-groups with greater than 25 procedures with complete PRO data). CORE acknowledges the complexity of these data and has assessed face validity from our TEP and Patient Working Group to capture stakeholder input to ensure the measure's scientific rigor over time.

1. Measure Introduction

1.1 Measure Overview

CMS has contracted with CORE to adapt claims-based hospital measures to assess the quality of care provided to Medicare beneficiaries undergoing elective primary THA/TKA by clinicians eligible to participate in the QPP MIPS. As part of this contract, CORE adapted CMS's existing hospital-level measure, the hospital-level THA/TKA PRO-PM, to assess individual or groups of MIPS participating clinicians. The re-specified measure will assess each clinician's improvement rate relative to that of other MIPS participating clinicians with similar patients.

1.2 Key Terminology

There exist many acronyms related to patient-reported outcomes. Throughout this report, we use the terminology advanced by the National Quality Forum (NQF): a "**PRO**" refers to the concept of a patient-reported outcome, a "**PROM**" refers to a survey instrument that captures patient-reported outcomes, and a "**PRO-PM**" is a performance measure that uses PRO data to define the measure outcome.[4]

Throughout this report, we use existing terminology specific to MIPS:

- **Eligible clinicians (ECs):** physicians, physician assistants, nurse practitioners, clinical nurse specialists, and certified registered nurse anesthetists who bill under Medicare Part B (81 FR 77036).[5] ECs are identified by a unique combination of Taxpayer Identification Number (TIN) and National Provider Identifier (NPI) numbers.
- **Clinician Groups (EC Group)**: ECs may participate as a single clinician, as a group (TIN with two or more clinicians), or as a virtual group (two or more TINs of solo practitioners and small groups of fewer than ten clinicians).

1.3 THA/TKA as a Measure of Quality

1.3.1 Importance

Elective THA/TKAs are most commonly performed for degenerative joint disease or osteoarthritis, which is the most common joint disorder in the US,[6] affecting more than 32.5 million, or 1 in every 7, US adults.[7] This condition is one of the leading causes of disability among non-institutionalized adults,[8] and roughly 80% of patients with osteoarthritis have some limitation in mobility.[9] Osteoarthritis also significantly burdens the health care system—in 2017, osteoarthritis was the second most expensive condition treated across all payers in US hospitals,[10] and in 2018, there were approximately 1,128,000 hospitalizations for osteoarthritis.[11, 12] Over 1 million THAs and TKAs are performed yearly in the US, of which 60% are paid for by Medicare.[13] THAs and TKAs offer significant improvement in quality of life by decreasing pain and improving function in a majority of patients, without conferring a high risk of complications and/or death.[14-17] As the goal of the procedures is to improve quality of life, THA and TKA are ideal candidates for assessing PROs.[18]

Due to their frequency and cost, THA and TKA are priority areas for outcome measure development. The number of THA and TKA procedures performed is expected to be over 2 million annually by 2030 and is expected to cost an estimated $50 billion in Medicare expenditures annually.[13]

Administrative claims-based elective primary THA/TKA risk-standardized complication and readmission measures at the hospital-level have been publicly reported since 2013, assessing outcomes important to patients and clinicians.[19] Neither of these measures, however, capture the reasons for which patients undergo elective THA and TKA: Will I have less pain and more mobility after the procedure? In short, will my quality of life be improved after undergoing the procedure? Therefore, a quality measure based upon PRO data provides both patients and providers with a unique and critical perspective on care.

THA/TKA procedures offer a particularly rich testbed for developing quality measures based upon patient-reported experiences and piloting performance measures based upon PROMs. Assessments of relative performance of improvement in PROs for clinician and/or clinician groups will provide targets to clinicians for efforts to improve quality of care.

1.3.2 Performance and Preventability

THA/TKA procedures are commonly performed in older patients who have marked pain and functional limitation preoperatively. These patients often experience significant improvements postoperatively. However, not all patients experience benefit from THA/TKA procedures, and many note their preoperative expectations for functional improvement were not met.[20-24] This variation in outcomes has been well documented in US clinical practice[25-27] and continues to be investigated in hospitals across the US and in the UK.[28, 29] Readmission and complication rates following THA/TKA vary across hospitals,[19, 30] and outcomes also vary significantly across hospitals worldwide—data collected in the United Kingdom (UK) indicate greater than 15% differences among hospitals in the proportion of patients who improved after the procedure.[28, 29]

There is also evidence of variation in performance at the clinician level, with risk-standardized complication rates (RSCRs) following THA/TKA across providers showing a median RSCR of 2.73% and an interquartile range of 2.39% - 3.16% for eligible clinicians, and a median RSCR of 2.75% and IQR of 2.48% - 3.09% for eligible clinician-groups.[31] Additionally, surgical approach (anterior vs posterior or lateral) has a significant impact on patients' outcomes such as risk of major surgical complications.[32]

1.3.3 Measurement Gap

This PRO-PM fills an important gap in the assessment of quality of care given to THA/TKA recipients. Other clinician-level PRO-PMs addressing total joint replacement exist, including several by Focus on Therapeutic Outcomes, Inc. (FOTO) and one developed by the Minnesota Community Measurement (MNCM) group in use in the state of Minnesota (NQF #2653). The FOTO measures, however, use proprietary software and focus on individual joints—not specifically on THA/TKA patients. NQF #2653 was developed for TKA recipients only, limiting the scope of the measure, and assesses an average change score for all eligible patients within an orthopedic practice. Specifically, we have heard stakeholder and patient concerns about averaging change in PROM scores across patients, which can make a measured entity (such as a clinician or clinician group) with all patients experiencing an average improvement appear the same as an entity where half of the patients do very well while the other half do very poorly. NQF #2653 also uses PROM surveys that the hospital-level THA/TKA PRO-PM development TEP previously considered but decided against using because they were either proprietary or did not offer the TEP's preferred information. In addition, CMS has developed procedure-specific process measures to incentivize the collection of PRO data (Functional Status Outcomes for Patients Receiving Primary Total Hip Replacements and Functional Status Outcomes for Patients Receiving Primary Total Knee Replacements; this PRO-PM will measure outcomes).

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.[18, 33-35] 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.[36]

1.3.4 Feasibility and Usability

Multiple initiatives expanding the use of PRO-PMs for THA/TKA procedures afford an opportunity to initiate public reporting within the US. There are several efforts led by orthopedic surgeons and their

professional societies to create regional and national patient registries. In addition, the Office of the National Coordinator for Health Information Technology (ONC) and CMS have included a process electronic clinical quality measure (eCQM) indicating collection of a THA/TKA PROM in the Promoting Interoperability (PI) programs and have developed eCQMs to promote THA/TKA PROM data collection for MIPS.[37] These initiatives are driven by an interest in improving clinical care and the need to evaluate long-term device safety, further prompted by recent publicized orthopedic device failures.[38]

In addition to the fact that the field of orthopedics is advanced in its development and use of validated PROMs for research, elective procedure such as THAs/TKAs provide a clear time zero (a reference time) for measurement: the date of the procedure. This allows for the use of a standardized measurement timeframe across clinicians and clinician groups.

In addition, Patient Working Group members noted strong support for the measure concept and recognized the importance of using PRO-PMs as tools for patients to help choose clinicians as well as an opportunity for clinicians to reflect on and improve their quality of care. Patients were supportive of both clinician-specific measure 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.

1.4 Measure Use

CORE is re-specifying the existing hospital-level THA/TKA PRO-PM for assessing individual or groups of MIPS participating clinicians. CMS's Meaningful Measures Framework supports alignment of quality measures across programs to minimize provider burden.[39] The re-specified measure will assess each clinician's improvement rate relative to that of other MIPS participating clinicians with similar patients.

1.5 Approach to Measure Development

CMS contracted with CORE to lead the re-specification of the hospital-level THA/TKA PRO-PM for use in MIPS under the guidance of CMS. The CORE Project Team consists of a multidisciplinary group of individuals with expertise in measure development, health services research, clinical medicine, statistics, and measurement methodology.

We developed this measure in consultation with national guidelines for publicly reported outcome measures, including the CMS Measure Management System Guidance,[40] those published by NQF,[41, 42] and as articulated in the American Heart Association scientific statement, "Standards for Statistical Models Used for Public Reporting of Health Outcomes."[43] Following these standards has ensured a transparent and comprehensive process with expert input throughout development (see Acknowledgements).

Below we review our approach to measure development.

1.5.1 Information Gathering

This measure leveraged the findings from the hospital-level THA/TKA PRO-PM's thorough literature reviews and environmental scan.[44] In preparation for measure development, the CORE hospital-level

THA/TKA PRO-PM measure development team conducted a literature review examining THA/TKA and PROM use to identify and define the technical decisions to be made in building this measure. The team also performed a systematic literature review of multivariable risk models predicting THA/TKA PROs to inform risk model development. They interviewed experts involved in implementing PROMs for quality assessment, including international experts experienced in national public reporting and US experts launching PROMs as part of THA/TKA registries.

1.5.2 Expert and Stakeholder Input

CORE consulted with a TEP, clinical orthopedic expert Dr. Kevin Bozic, a Clinical Working Group, and a Patient Working Group for input on the re-specified measure specifications (see Appendix A). The Patient Working Group supported the re-specification of the hospital-level THA/TKA PRO-PM and suggested additional topics they felt should be considered in the measure. The TEP and Clinical Working Group shared critical feedback regarding the procedures captured in the cohort, PROM data collection timeframe, and the risk model.

2. Methods

2.1 Overview

This MIPS THA/TKA PRO-PM aligns with the existing hospital-level THA/TKA PRO-PM specifications, including PROMs, PROM assessment timeframes, cohort, outcome, and risk adjustment, and with the MIPS THA/TKA Complication measure's clinician attribution algorithm, with one refinement. We consulted with our orthopedic clinical expert, TEP, Clinical Working Group, and Patient Working Group throughout the measure re-specification process.

2.2 Data Sources

The MIPS THA/TKA PRO-PM primarily uses PRO and risk variable data collected and submitted by clinicians or clinician groups, matched per procedure to CMS administrative claims data. PRO and additional risk variable data are collected preoperatively, and PRO data are collected again postoperatively on patients undergoing an elective primary THA/TKA. (For purposes of measure development and testing, we utilized PRO and risk variable data collected by hospitals participating in the voluntary PRO data collection effort in CJR.) CMS administrative claims data are used to identify eligible elective primary THA/TKA procedures for the measure cohort and a majority of risk variables, including patient demographics and clinical comorbidities identified up to 12 months prior to the procedure. Medicare Part A inpatient data and Medicare Part B inpatient claim and claim line data from the IDR are used to match patients to clinicians and clinician groups who billed for the procedure. Three additional data sources provide data for the measure as follows: the Medicare EDB identifies Medicare FFS enrollment and race, the MBSF allows for determination of dual eligibility status, and the American Community Survey data allow for derivation of the AHRQ SES index score. Data from these sources are linked for patients undergoing elective primary THA and TKA procedures for the measurement period. Patients with complete preoperative and postoperative PRO and risk variable data were included in the dataset used for development and testing of this measure.

2.2.1 Limitations

In the absence of routinely collected prospective PRO data at the national level, we utilized existing PRO and risk variable data voluntarily collected by hospitals participating in the CJR model. Although the CJR data reflect a voluntary data sample incentivized at the hospital-level, they contain the appropriate cohort, risk variable, and linking data needed for our measure testing. We recommend ongoing reevaluation of the measure specifications in broader datasets over time.

2.2.2 Missing Data and Non-Response

For measure development and testing, only elective primary THA/TKA procedures with complete PRO data were included. Complete PRO and risk variable data were defined as the submission of preoperative PROM and risk variable data with no missing or out-of-range values for required data elements 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.

Due to the voluntary nature of PROs, PRO data were not collected from all patients undergoing primary elective THA/TKA procedures during the measurement period. We understand that accounting for potential non-response bias is important for this measure, as poorly or incompletely collected data may be asymmetrically distributed across lower socioeconomic or disadvantaged populations, which may directly affect measure scores. We examined the impact and addressed potential bias resulting from patients not included in the measure due to missing PRO data (patients with incomplete PRO data) and non-response (patients for which no PRO data were submitted) by using stabilized inverse probability weighting, which aligns with the approach utilized by the hospital-level THA/TKA PRO-PM (see Section 2.7.1 Response Bias).

Response rates for PRO data for this measure will be calculated as the percentage of elective primary THA or TKA procedures for which complete and matched preoperative and postoperative PRO data have been submitted, divided by the total number of eligible THA or TKA procedures performed by each clinician or clinician group.

2.2.3 Generalizability

The CJR data used for testing were incentivized at the hospital level, with response rates 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.

2.3 Measure Cohort

The cohort (target population) includes Medicare FFS patients 65 years of age and older undergoing elective primary THA or TKA procedures, not including patients with hip and/or pelvic fractures and/or revision THAs/TKAs as detailed below. The cohort definition aligns with the existing hospital-level THA/TKA PRO-PM.

2.3.1 Inclusion Criteria

- 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.
- Aged 65 or older.
- Discharged alive from a non-federal short-term acute care hospital.
- Undergoing only elective primary THA/TKA procedures (not including patients with fractures of the pelvis or lower limbs, concurrent partial hip or knee arthroplasty procedure, concurrent revision, resurfacing, or implanted device/prosthesis removal procedure, mechanical complication coded in the principal discharge diagnosis field on the index claim, malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field on the index claim, or a transfer from another acute care facility for the THA/TKA).

2.3.2 Exclusion Criteria

- 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.
- Patients who die within 300 days of the procedure are excluded. Patients who die within 300 days are unable to complete PROM data in alignment within the postoperative PROM collection timeframe. (In testing this PRO-PM with CJR data, patients who died within 270 days were excluded from the cohort to align with the postoperative window for CJR voluntary data collection.)
- Patients who leave against medical advice.

In addition, patients with more than two THA or TKA procedure codes on their index hospitalization claim are removed from the cohort. The occurrence of more than two THA or TKA procedure codes is likely a coding error.

2.4 Measure Outcome

Patient improvement will be measured using the joint-specific instruments selected for the hospital-level THA/TKA PRO-PM: the HOOS, JR for THA patients and the KOOS, JR for TKA patients. The preoperative data collection timeframe will be 90 to zero days before the procedure and the postoperative data collection timeframe will be 300 to 425 days following the procedure. (The measure was developed and tested with postoperative data collected 270 to 365 days after the procedure, as per PRO data collection in the CJR model. Extensive input from clinical experts following measure development, however, strongly recommended a refinement to the postoperative data collection period to better align with clinical workflow and typical one-year follow-up scheduling, and to allow for better postoperative PRO data capture. Based on this input, we propose measure specifications with a postoperative PRO data collection period of 300 to 425 days after the procedure. We anticipate, based

upon extensive stakeholder input, this will result in limited impact to the measure's scientific acceptability while significantly increasing clinical acceptance and response rates.)

The patient-level improvement outcome is defined as a binary outcome (yes/no) of meeting or exceeding substantial improvement, SCB. The SCB threshold is a fixed threshold identified by the developers of the HOOS, JR and KOOS, JR, who used an anchor-based patient-reported satisfaction question to identify the amount of change in PROM scores found among patients reporting "great improvement" following a THA or TKA.[45] The clinician and clinician-group outcome is a risk-adjusted proportion of elective primary THA/TKA patients who meet SCB improvement. The risk adjustment model accounts for case mix differences and aligns with the hospital-level THA/TKA PRO-PM's risk model. The measure will estimate a clinician- and clinician group-level RSIR PROs following elective primary THA and/or TKA for Medicare FFS patients 65 years of age and older. The clinician- and clinician group-level RSIRs will account for within and between clinician/clinician group case mix and represent measures of quality of care following elective primary THA and TKA.

2.5 Attribution

The MIPS THA/TKA PRO-PM uses the clinician attribution methodology developed by the MIPS THA/TKA Complication measure team, with one refinement. For this measure, we narrowed the list of clinician specialties primarily responsible for the orthopedic surgery to five specialties. The measure attributes the outcome for each patient in the cohort to a single clinician. Each patient is attributed to the clinician who bills for the Part B Physician/Supplier claim for the THA or TKA procedure (hereafter referred to as the Billing Surgeon) during the index admission. Conceptually, this is the clinician with the primary responsibility for the procedure and procedure-related care. In practice, however, patients may have multiple claims from different clinicians. In order to resolve any ambiguities, the Billing Surgeon is assigned through an algorithm (Appendix C Figure C1) as described below.

The algorithm uses billing claims to identify clinician(s) who bill for a THA (Current Procedural Terminology [CPT®] code 27130) or TKA (CPT®® code 27447 or CPT® code 27446 if CPT® code 27447 not billed) (steps 1-3 below). These CPT® codes allow identification of responsible surgeons in alignment with the THA and/or TKA procedures included in the measure cohort.

1. If only one clinician bills for a THA (CPT® code 27130) or TKA (CPT® code 27446 or 27447) for a patient, the algorithm identifies and assigns this individual as the Billing Surgeon.
2. If two or more clinicians bill for THA/TKA procedures (CPT® 27130, 27447, or 27446), the algorithm seeks to identify a 'key' physician among them.
    a. The algorithm identifies and excludes assignment to clinicians who were assistants-at-surgery (assistant surgeon with CPT® modifier 80 or 82, minimum assistant surgeon with CPT® modifier 81). In this step, the algorithm assigns the Billing Surgeon as the clinician who billed for a THA or TKA procedure and is not an assistant-at-surgery.
    b. If two or more clinicians remain after the exclusion of assistants-at-surgery, then the algorithm identifies whether there is a single clinician who was an orthopedic surgeon (Medicare Specialty Code 20) and assigns this as the Billing Surgeon.

3. If the algorithm cannot identify a Billing Surgeon, it identifies whether an Operator is listed on the institutional claim. The algorithm then defaults assignment to the Operator listed on the institutional claim. To identify the unique TIN/NPI combination for the Operator, the Operator's NPI is matched to the TIN with the most Part B allowed charges during the index admission or during the measurement year if the eligible clinician did not bill during the index admission.

Finally, if a Billing Surgeon or Operator cannot be identified with the steps above, the patient is not assigned to a Billing Surgeon and is excluded from the measure.

During the development of the MIPS THA/TKA Complication measure, clinical experts and the MIPS TEP supported attribution to the Billing Surgeon using this algorithm.

During measure testing of the MIPS THA/TKA PRO-PM, we identified that the original attribution methodology assigned a few patients to clinicians not primarily responsible for the THA/TKA procedures, such as Physician Assistants or clinicians with specialty descriptions outside of orthopedics (such as obstetrics/gynecology). Although this did not happen often (N = 12, 0.94% of measure cohort), the patient assignment was incongruous with the intention of the attribution methodology. We, therefore, propose the measure attribution only consider the following clinician specialties: orthopedic surgery, sports medicine, hand surgery, osteopathic manipulative medicine, and general surgery. Specifically, we recommend first limiting the Part B claim lines to those with the above clinician specialties and applying the original attribution algorithm afterwards. This refinement will help increase the face validity of the attribution methodology. Our team's orthopedic expert and TEP agreed with the refinement to the attribution methodology.

2.6 Risk Adjustment

Risk adjustment is designed to account for patient-level factors that are clinically relevant, have strong relationships with the outcome, and are outside of the control of the reporting entity without obscuring important quality differences. Risk factors can increase (or decrease) the likelihood that a patient experiences a certain outcome. Risk adjustment puts measured entities on a level playing field when comparing performance across providers.

The MIPS THA/TKA PRO-PM utilizes the risk model finalized for the hospital-level THA/TKA PRO-PM. The hospital-level THA/TKA PRO-PM development team used a consensus-based approach to identify and vet clinically relevant risk variables important in predicting the improvement outcome, including a systematic literature review and environmental scan, a survey of orthopedists, consultation with an expert clinical consultant, extensive input from the TEP, and detailed public comments. This team tested the risk model during measure development and gained stakeholder support from clinical experts and patients on the approach and testing results. We will use this risk model because patient-level risk prediction should not be impacted by attribution and to align with the hospital-level THA/TKA PRO-PM specifications.

Of note, the final risk variables in the risk model come from two sources:

1. Risk variables identified by THA/TKA professional societies and public comment, finalized in the CJR final rule, and submitted by CJR participant hospitals, and
2. Clinical comorbidities identified in Medicare claims data for the 12 months prior to THA/TKA procedures. The Condition Category (CC) groupings below reflect groupings of The International Classification of Disease ICD-10-CM diagnosis codes into clinically relevant categories, from the Hierarchical Condition Category (HCC) system.[46, 47] CMS uses modified groupings, but not the hierarchical logic of the system, to create risk variables.

The final risk variables are as follows:

- Age, in years
- Male sex
- Body Mass Index (BMI), in kg per m^2
- Health literacy (assessed by response to Single Item Literacy Screener questionnaire, "Comfort Filling Out Medical Forms by Yourself")
- Back pain at preoperative assessment (Quantified Spinal Pain: Patient-reported Back Pain, Oswestry Disability Index question)[48, 49]
- Pain in non-operative lower extremity joint (Total painful joint count: Patient-reported in Non-operative Lower Extremity Joint)[49]
- Narcotic use for ≥90 days
- Baseline Patient-reported Outcomes Measurement Information System (PROMIS) Global Mental Health subscale score
- Severe infection; other infectious diseases (CC 1, 3-7)
- Liver disease (CC 27-31)
- Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)
- 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)

Detailed information about the development of this risk model approach and testing can be found in Sections 2.6 and 3.3 of the Hospital-level THA/TKA PRO-PM Measure Methodology Report Version 1.0.

2.6.1 Social Risk Factors

During development of the hospital-level THA/TKA PRO-PM, the measure development team focused on developing a parsimonious model that included clinically relevant variables strongly associated with achieving SCB following an index procedure. The team used a two-stage approach, first identifying the comorbidity or clinical status risk factors that were most important in predicting the outcome then

considering the potential addition of social risk factors. The hospital-level THA/TKA PRO-PM development team's patient and technical experts strongly supported including health literacy, which reflects social risk, in the risk model for a PRO-based measure, due to the nature of PRO data requiring patients to complete survey instruments as part of measurement. For this reason, the team included it in the candidate risk variable list during the initial stage of risk variable selection and in the final risk model.

Additional social risk factors were examined following initial risk model development. Similar to the hospital-level THA/TKA PRO-PM, we tested associations of dual eligibility and the AHRQ SES index lowest quartile (low SES) with SCB improvement following elective primary THA/TKA to explore the impact of social risk on the measure outcome. Likewise, we included race in the social risk factor analyses based upon literature specifically documenting racial and ethnic disparities in THA/TKA offer and acceptance rates as well as outcomes.[50, 51]

We examined the associations of dual eligibility, AHRQ SES index lowest quartile (low SES), and race with the measure outcome using the Development Dataset with bivariate and multivariate analyses. Bivariate and multivariate analyses showed no statistically significant associations between non-white race and SCB improvement, AHRQ SES index lowest quartile and SCB improvement, or dual eligibility and SCB improvement (see Section 3.3.1 Table 11). In addition, accounting for these risk factors in the model had little effect on clinician- and clinician group-level RSIRs. When RSIRs calculated with no social risk factors in the risk model were compared to RSIRs calculated with each of the three risk factors included individually and together in the risk model, correlation coefficients indicated near-perfect correlation in our data (see Section 3.3.1 Table 12).

Additional analyses of clinician and clinician group proportion of dual eligible, AHRQ SES index lowest quartile, and non-white race patients by RSIRs (see Section 3.3.1 Figures 4 through 9) indicate that clinicians and clinician groups with the lowest proportion of dual eligible patients and those clinicians and clinician groups with the highest proportion of dual eligible patients have similar RSIR distributions. These data do not provide evidence of significant differences in RSIRs due to the proportion of clinicians' or clinician groups' patients with dual eligibility, AHRQ SES index lowest quartile, or non-white race.

Given these results, we have not included dual eligibility, AHRQ SES index, or non-white race in the risk model(s) at this time. However, given the associations between social risk factors and response in our data and supported by published literature,[52-54] we have included these social risk variables in the statistical approach to addressing non-response bias (see Section 2.7.1 below). In addition, due to existing disparities in access to and outcomes for THA/TKA, the TEP, the Patient Working Group, and other stakeholders strongly urged CMS to consider accounting for social risk in measure implementation (such as through stratification) to avoid unintended consequences. We will continue to assess the impact of social risk on this measure moving forward.

2.7 Statistical Approach to Model Development and Testing

The number of patients with complete PRO data for an elective primary THA or TKA procedure (excluding patients with staged elective primary THA/TKA procedures during the measurement period) was 19,429. These data were randomly divided 60%/40% into a Development Dataset (N=11,653) and a Validation Dataset (N=7,776). The Development Dataset was used for risk model development, and the Validation Dataset was used for risk model validation.

In the building of risk models for claims-based measures, CORE has previously used the strength of association between the risk variable and the measure outcome to empirically guide risk variable selection. When expert input deems it appropriate, we force in additional risk variables, such as those that indicate frailty, that might have an important influence on the measure outcome and yet might not be selected for the model based purely on statistical considerations. In this way, CORE's risk models have always reflected both empirical data and clinical input. This approach has produced robust risk models that have been repeatedly and successfully validated against medical record data.

For the hospital-level THA/TKA PRO-PM, the team applied the same principles but recognized that PRO-PM development, particularly when based upon a voluntary data sample, may require a greater reliance on clinical input to select risk variables than traditional claims-based outcome measures. For the risk model, the hospital-level THA/TKA PRO-PM development team identified candidate risk variable selection based on empirical findings through a thorough literature review, exploration of data-driven risk factors, and iterative TEP and expert input including a survey of the TEP that asked individuals to rank the importance and feasibility of clinical variables for use in a PRO-PM risk model. Input from clinical consultants was used to finalize a list of clinically relevant and important risk variables for risk adjustment of a THA/TKA PRO-PM.

Using the risk model developed by the hospital-level THA/TKA PRO-PM development team, we assessed model performance examining the model performance (C-statistics), model calibration (lack of fit), model discrimination in terms of predictivity (range of observed outcomes among deciles of predicted outcomes), and distribution of model residuals. We calculated the model estimates as well as the coefficients and 95% Confidence Intervals (CIs) for risk-adjustment variables in the Development and Validation Datasets.

For risk model development, we first fit a logistic regression model with the selected covariates (risk variables). We then 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 (see Section 2.8 Calculation of Measure Score).

2.7.1 Response Bias

Due to the voluntary nature and the burden of novel data collection for PRO data, we understand that accounting for potential non-response bias is important for this measure. Also, the fact that poorly or incompletely collected data may be asymmetrically distributed across lower socioeconomic or disadvantaged populations may directly affect measure scores. The hospital-level THA/TKA PRO-PM

development 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, they decided to address potential response bias using stabilized inverse probability weighting, as it would not modify the risk model, and would not assume the form of a relationship between submission score and outcome (as suggested by Garrido[55] and by Thoemmes and Ong[56]). 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 (the source of the PRO data for development and testing) submitting complete PRO data during the measurement period among 1,254 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:
   a) Procedures for which complete PRO and risk variable preoperative data and complete PRO postoperative data were submitted ("complete PRO submission," N = 19,429).
   b) Procedures for which incomplete PRO and risk variable data were submitted (including submissions with missing data elements and submissions of only preoperative PRO data or only postoperative PRO data ("incomplete PRO submission," N = 17,220).
   c) 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 statistical differences in the case mix.
4) The hospital-level THA/TKA PRO-PM development 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.[52-54, 57] 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 Weights (IPW) were calculated for each of the three groups. For the complete responders, the stabilized weights were calculated using the following formula: $\frac{P(Z=1)}{P(Z=1|x)}$ where $(Z = 1)$ represents the complete responders. Stabilized weights produce estimates with smaller variance and less extreme values compared to using the standard non-stabilized weights calculated in the following way: $\frac{1}{P(Z=1|x)}$. Section 3.8 Table 18 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 the calculation of the risk-adjusted and bias-adjusted RSIRs.

Incorporating the stabilized weights in the calculation of the RSIRs helps to reduce non-response bias 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.

2.8 Calculation of Measure Score

The MIPS THA/TKA PRO-PM is a clinician- and clinician group-level measure that will be calculated and presented as an RSIR, producing a performance measure per clinician or clinician group that accounts for patient case mix and applies stabilized IPW to address potential non-response bias. The RSIR represents a measure of quality of care following primary elective THA and TKA.

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.

For the hierarchical logistic regression model: Let Y_{ij} denote the outcome (equal to one if the patient meets the SCB threshold, zero otherwise) for patient *i* attributed to a clinician or clinician group *j*; $\boldsymbol{Z}_{ij}$ denotes a set of risk factors for patient i attributed to a clinician or clinician group j; and n_j is the number of index admissions attributed to the clinician or clinician group j. We assume the outcome is related linearly to the covariates via a logit function:

$$logit\big(\text{Prob}(Y_{ij}=1)\big) = \alpha_j + \boldsymbol{\beta Z}_{ij}$$

where $\alpha_j = \mu + \omega_j;\ \omega_j \sim N(0, \tau^2)$

Where $\alpha_j$represents the clinician- or clinician group-specific intercept, μ is the adjusted average intercept over all clinicians or clinician groups in the sample, ω_j 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. Random effects account for the assumption that underlying differences in quality of care across clinicians and clinician groups lead to systematic differences in patient outcomes.

To account for potential response bias, we calculated stabilized 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 Section 2.7.1 for a detailed description of the analytic approach to addressing potential response bias). Next, we fit the hierarchical logistic

regression model to the corresponding parameters along with the stabilized IPW to adjust for response bias when we ran the model.

From the hierarchical model, we calculated the clinician- and clinician group-specific RSIRs as the ratio of a clinician's or clinician group's "predicted" number of improvements to their "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 patient characteristics and the clinician- or clinician group-specific intercept.

2.9 Measure Testing

The dataset for measure development and testing included data from CJR participant hospitals that submitted complete preoperative and postoperative PRO and risk variable data for at least one elective primary THA/TKA procedure performed from July 1, 2016 through June 30, 2018. Complete PRO and risk variable data were defined as the submission of preoperative PROM and risk variable data with no missing or out-of-range values for required data elements that could be matched to postoperative PROM data with no missing or out-of-range values for an elective primary THA/TKA procedure identified in claims data for the measurement period. The number of patients with complete PRO data for an elective primary THA or TKA procedure (excluding patients with staged elective primary THA/TKA procedures during the measurement period, which is defined as two or more procedures performed during separate inpatient admissions) was 19,429. These data were randomly divided 60%/40% into a Development Dataset (N=11,653) and a Validation Dataset (N=7,776). A Combined Dataset including only those clinicians or clinician groups with at least 25 THA and TKA patients with complete PRO data submission was used for reliability and validity testing.

2.9.1 Model Performance Testing

We assessed the risk model in the Development Dataset by examining the model performance (C-statistics), model calibration (lack of fit), model discrimination in terms of predictive ability (range of observed outcomes among deciles of predicted outcomes), and distribution of model residuals. We calculated the model estimates as well as the coefficients and 95% CIs for risk-adjustment variables for the entire Development Dataset.

To assess model performance, we computed discrimination and calibration statistics for assessing model performance[58] for the risk model, 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 is able to distinguish between a patient with and without an outcome).

2. Predictive ability (discrimination in predictive ability measures the ability to distinguish high-risk subjects from low-risk subjects; good discrimination is 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).

2.9.2 Measure Results Testing

Meaningful differences in performance measure scores are assessed by calculating the distribution of clinician- and clinician group-level RSIRs. Variation in RSIRs indicates a clinically meaningful quality gap in the delivery of care to patients undergoing primary elective THA/TKA, as some clinicians and clinician groups may 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).[59] The MOR represents the median increase in odds of the patient outcome (an 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 (N=1254 clinicians and N=526 clinician groups in the total dataset), always comparing the higher-performing clinicians or clinician groups to the lower performing clinicians or clinician groups. The MOR is interpreted as a traditional odds ratio (OR) would be.

2.9.3 Reliability Testing

2.9.3a Data Element Reliability Testing

Data element reliability of the PROM data used in the calculation of this measure was assessed by reliability testing conducted during the development and validation of the joint-specific PROMs on which this MIPS THA/TKA PRO-PM is based.

HOOS, JR Reliability

Internal consistency: The developers of the HOOS, JR[2] assessed internal consistency reliability of using the Person Separation Index (PSI). The PSI was used in two data samples, the Hospital for Special Surgery (HSS) cohort and the Function and Outcomes Research for Comparative Effectiveness in Total Joint Replacement (FORCE-TJR), a nationally representative joint replacement registry. A higher value on the PSI indicates greater ability to differentiate patients with varying levels of ability, which provides evidence of good internal consistency. For testing internal consistency for the HOOS, JR, a PSI value greater than 0.7 was considered acceptable.[2] 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 HOOS in several validation studies.[60-63] 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 KOOS, JR[3] assessed internal consistency reliability of 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 a greater ability to differentiate patients with varying levels of ability, which provides evidence of good internal consistency. For testing internal consistency for the KOOS, JR, a PSI value greater than 0.7 was considered acceptable.[3] 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 KOOS.[64] To examine test-retest reliability, the KOOS was administered to patients twice prior to the procedure 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.

Data element reliability of the codes identified through Medicare claims data used to define the cohort and for risk adjustment is expected as a result of the routine auditing of these billing data by CMS. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud and audits important data fields used in our measures, including diagnosis and procedure codes, and other elements that are consequential to payment.

Likewise, the risk variables collected with PRO data that will be included in the risk model were defined during the hospital-level THA/TKA PRO-PM development, which spanned over several years through multiple, iterative steps that incorporated stakeholder input on feasibility, clinical capture, accuracy, reproducibility, and clinical face validity. These steps included: surveying orthopedic practices regarding the feasibility, uniformity, and reliability of risk variables identified by clinical experts and published literature; attending a consensus summit by orthopedic specialty societies to narrow and prioritize clinical risk variables for prospective collection as part of the CJR model (these recommendations were adopted *in toto* by CMS); seeking additional clinical and empirical evaluation of CJR data; and receiving TEP approval.

2.9.3b Measure Score Reliability Testing

We performed reliability testing at both the clinician and the clinician group level. We calculated the average signal-to-noise ratio based on the weighted average of measured clinicians or clinician groups with at least 25 procedures with complete PRO data. We also examined measure score reliability across a range of minimum case volume thresholds to support CMS's implementation planning.

2.9.4 Validity Testing

2.9.4a Data Element Validity Testing

Data element validity is ascertained through validity testing conducted during the development and testing of the joint-specific PROMs on which this MIPS THA/TKA PRO-PM is based. All validity testing for the HOOS, JR and KOOS, JR instruments was conducted by the PROM developers.[2, 3]

HOOS, JR Validity

Responsiveness: Responsiveness of the HOOS, JR to changes following a THA was evaluated using standardized response means, and then examined against other previously validated PROMs (HOOS domains, Western Ontario and McMaster Universities Osteoarthritis Index [WOMAC] domains) in the HSS cohort and the FORCE-TJR registry at two years after a THA procedure.[2] A standardized response mean greater than 0.8 was considered large.[65]

External Validity: External construct validity was evaluated using Spearman's correlations between the HOOS, JR, the HOOS, and the WOMAC. A Spearman's correlation coefficient of 0.8 or greater was considered very high external validity.[66] External correlations were assessed using a scatterplot overlying a contour plot based on bivariate kernel density estimation between the HOOS, JR and HOOS domains.[2]

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.[2]

KOOS, JR Validity

Responsiveness: Responsiveness of the KOOS, JR to changes following TKA was evaluated using standardized response means, and then examined against other validated PROMs (KOOS domains, WOMAC domains) in the validation cohort.[3] A standardized response mean greater than 0.8 was considered large.[65]

External Validity: External construct validity was evaluated using Spearman's correlations between the KOOS, JR, the KOOS, and the WOMAC. A Spearman's correlation coefficient of 0.8 or greater was considered very high external validity.[66] External correlations were assessed using a scatterplot overlying a contour plot based on bivariate kernel density estimation between the KOOS, JR and KOOS domains.[3]

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.[3]

2.9.5 Face Validity Testing

We have re-specified this measure in consultation with national guidelines for publicly reported outcome measures, with outside experts, and with the public. We assessed face validity by asking 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):

- Statement #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.
- Statement #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.

3. Results

3.1 Measure Cohort

Characteristics of the 11,653 patients in the Development Dataset and the 7,776 patients in the Validation Dataset are presented in Table 1. The cohort flowchart can be found in Figure 1.

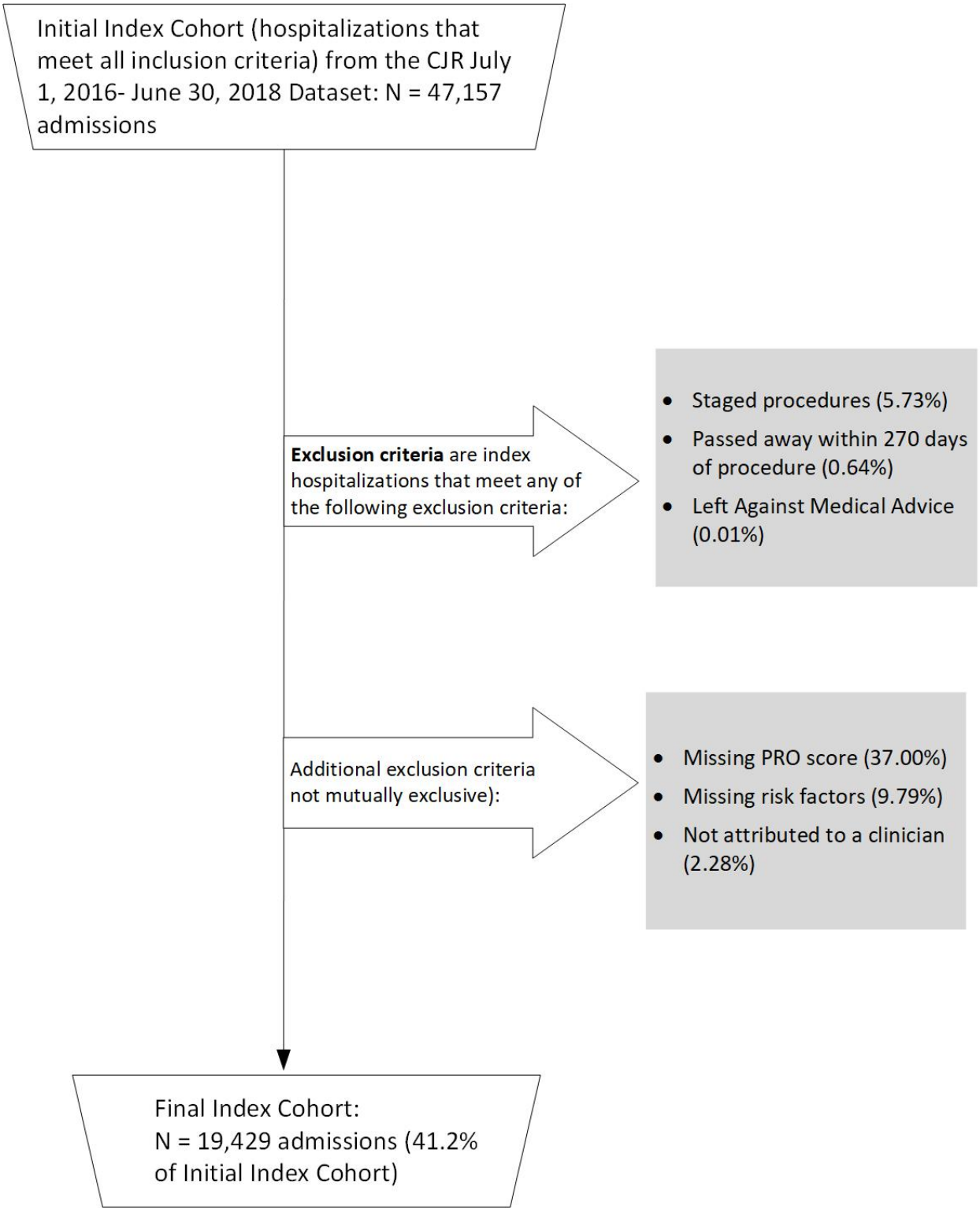
Table 1. Patient Characteristics in Development and Validation Datasets

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%)

Characteristics		Development Dataset, N (%)	Validation Dataset, N (%)
	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%)
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	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%)
	Unknown	190 (1.63%)	125 (1.61%)

*Note: Missing AHRQ SES Index information in Development Dataset=27 (0.23%) and Validation Dataset=14 (0.18%)

Figure 1. Cohort Flowchart



3.2 Attribution

Characteristics of clinicians and clinician groups in the Development Dataset and Validation Dataset as well as the full sample are presented in Table 2 and Table 3, respectively. The distribution of procedure volumes for clinicians and clinician groups with ≥25 complete PRO data are presented in Table 4.

Table 2. 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 3. 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 4. Distribution of Volumes for Clinicians and Clinician Groups with ≥25 Procedures 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

3.3 Risk Model Development and Testing

The risk model results using the final combined dataset can be found in Table 5, including the frequency of risk variables and ORs. 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 an SCB.

Table 5. Final Risk Model Variables and Adjusted Odds Ratios (Logistic Regression Model): Development Dataset (Patient N = 11,653)

Risk Factors		Frequency	Odds Ratio (95% CI)
Age		-	1.00 (1.00–1.01)
Male		4,405 (37.80%)	0.82 (0.75–0.89)
BMI		-	1.01 (1.00–1.02)
Index admissions with an elective THA procedure		4,193 (35.98%)	1.38 (1.24–1.47)
Number of procedures (two vs. one)		67 (0.57%)	1.50 (0.84–2.70)
Mental Health Score Mean		-	0.99 (0.98–0.99)
Health Literacy	Not at all	2,015 (17.29%)	Reference
	A little bit	881 (7.56%)	1.27 (1.08–1.50)
	Somewhat	1,291 (11.08%)	1.59 (1.37–1.85)
	Quite a bit	2,079 (17.84%)	1.77 (1.55–2.02)
	Extremely	5,387 (46.23%)	1.98 (1.78–2.21)
Other Joint Pain	None	4,057 (34.82%)	Reference
	Mild	2,897 (24.86%)	0.92 (0.83–1.02)
	Moderate	2,890 (24.80%)	0.99 (0.89–1.10)
	Severe	1,470 (12.61%)	1.42 (1.24–1.63)
	Extreme	339 (2.91%)	1.90 (1.43–2.52)
Back Pain	None	4,459 (38.26%)	Reference
	Very Mild	2,905 (24.93%)	0.91 (0.82–1.01)
	Moderate	2,964 (25.44%)	0.86 (0.78–0.96)

Risk Factors		Frequency	Odds Ratio (95% CI)
	Fairly Severe	948 (8.14%)	0.95 (0.81–1.11)
	Very or Worst Severe	377 (3.24%)	1.46 (1.12–1.90)
Use of Chronic (≥ 90 days) Narcotics		2,032 (17.44%)	0.96 (0.87–1.07)
Severe infection; other infectious diseases (CC 1, 3–7)		2,023 (17.36%)	0.88 (0.79–0.98)
Liver disease (CC 27–31)		491 (4.21%)	0.81 (0.67–0.98)
Diabetes mellitus (DM) or DM complications (CC 17–19, 122–123)		3,013 (25.86%)	1.00 (0.91–1.10)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease (CC 40)		1,249 (10.72%)	0.92 (0.81–1.05)
Depression (CC 61)		1,832 (15.72%)	0.85 (0.76–0.96)
Other Psychiatric Disorders (CC 63)		1,839 (15.78%)	0.97 (0.86–1.09)
Coronary atherosclerosis or angina (CC 88–89)		2,878 (24.70%)	0.87 (0.79–0.96)
Vascular or circulatory disease (CC 106–109)		2,256 (19.36%)	0.97 (0.87–1.07)
Renal failure (CC 135–140)		1,637 (14.05%)	1.01 (0.90–1.13)

Table 6. Final Risk Model Variables and Adjusted Odds Ratios (Logistic Regression Model): Validation Dataset (Patient N = 7,776)

Risk Factors		Frequency	Odds Ratio (95% CI)
Age		-	1.00 (0.99–1.01)
Male		2,889 (37.15%)	0.82 (0.74–0.91)
BMI		-	1.00 (0.99–1.01)
Index admissions with an elective THA procedure		2,778 (35.73%)	1.39 (1.25–1.54)
Number of procedures (two vs. one)		49 (0.63%)	3.44 (1.44–8.20)
Mental Health Score Mean		-	0.99 (0.98–0.99)
Health Literacy	Not at all	1,267 (16.29%)	Reference
	A little bit	621 (7.99%)	1.22 (0.99–1.49)

Risk Factors		Frequency	Odds Ratio (95% CI)
	Somewhat	833 (10.71%)	1.70 (1.41–2.06)
	Quite a bit	1,410 (18.13%)	1.70 (1.44–2.00)
	Extremely	3,645 (46.88%)	1.95 (1.70–2.23)
Other Joint Pain	None	2,637 (33.91%)	Reference
	Mild	1,871 (24.06%)	0.81 (0.71–0.92)
	Moderate	2,007 (25.81%)	0.93 (0.82–1.06)
	Severe	1,046 (13.45%)	1.37 (1.16–1.63)
	Extreme	215 (2.76%)	2.16 (1.49–3.13)
Back Pain	None	2,869 (36.90%)	Reference
	Very Mild	1,979 (25.45%)	1.00 (0.89–1.14)
	Moderate	2,024 (26.03%)	1.03 (0.90–1.17)
	Fairly Severe	653 (8.40%)	0.95 (0.79–1.16)
	Very or Worst Severe	251 (3.23%)	1.52 (1.10–2.09)
Use of Chronic (≥ 90 days) Narcotics		1,358 (17.46%)	0.90 (0.79–1.03)
Severe infection; other infectious diseases (CC 1, 3–7)		1,386 (17.82%)	0.94 (0.83–1.07)
Liver disease (CC 27–31)		322 (4.14%)	0.90 (0.83–1.34)
Diabetes mellitus (DM) or DM complications (CC 17–19, 122–123)		2,005 (25.78%)	0.91 (0.70–1.14)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease (CC 40)		834 (10.73%)	0.95 (0.81–1.11)
Depression (CC 61)		1,180 (15.17%)	1.04 (0.90–1.21)
Other Psychiatric Disorders (CC 63)		1,260 (16.20%)	0.88 (0.76–1.02)
Coronary atherosclerosis or angina (CC 88–89)		1,872 (24.07%)	0.95 (0.85–1.07)
Vascular or circulatory disease (CC 106–109)		1,471 (18.92%)	0.82 (0.72–0.93)

Risk Factors	Frequency	Odds Ratio (95% CI)
Renal failure (CC 135–140)	1,116 (14.35%)	1.10 (0.95–1.27)

Table 7. 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		-	1.00 (1.00-1.01)
Male		7,294 (37.54%)	0.82 (0.76-0.87)
BMI		-	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		-	0.99 (0.98-0.99)
Health Literacy	Not at all	3,282 (16.89%)	Reference
	A little bit	1502 (7.73%)	1.25 (1.10-1.42)
	Somewhat	2,124 (10.93%)	1.63 (1.45-1.84)
	Quite a bit	3,489 (17.96%)	1.74 (1.57-1.93)
	Extremely	9,032 (46.49%)	1.97 (1.81-2.14)
Other Joint Pain	None	6,694 (34.45%)	Reference
	Mild	4,768 (24.54%)	0.88 (0.81-0.95)
	Moderate	4,897 (25.20%)	0.97 (0.89-1.05)
	Severe	2,516 (12.95%)	1.41 (1.26-1.57)
	Extreme	554 (2.85%)	2.00 (1.60-2.50)
Back Pain	None	7,328 (37.72%)	Reference
	Very Mild	4,884 (25.14%)	0.95 (0.88-1.03)

Risk Factors		Frequency	Odds Ratio (95% CI)
	Moderate	4,988 (25.67%)	0.93 (0.85-1.00)
	Fairly Severe	1,601 (8.24%)	0.95 (0.84-1.07)
	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%)	0.90 (0.83-0.98)
Liver disease (CC 27–31)		813 (4.18%)	0.85 (0.73-0.98)
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%)	0.90 (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)

Model performance statistics for the risk model for meeting or exceeding the SCB improvement threshold are provided in Table 8. In the Development Dataset, the C-statistic for the risk model is 0.61 and the predictive ability from the lowest to highest decile is 48.67% - 80.03%. In the Validation Dataset, the C-statistic for the risk model is 0.61 and the predictive ability from the lowest to highest decile is 52.44%- 81.14%. The *calibration indices* (γ0, γ1) used to assess the risk model for meeting or exceeding SCB improvement for the Validation Dataset are (0.02, 0.97). Table 9 shows there is not a significant lack of fit.

Results demonstrate the risk-adjustment model moderately controls for differences in patient characteristics. The C-statistics indicate adequate model discrimination across the 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.

The calibration values which are consistently close to zero at one end and close to one at the other end indicate 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 one, there is potential evidence of over-fitting. The calibration values of close to zero at one end and close to one on the other end indicate good calibration of the model between the Development and Validation Datasets.

In the risk decile plots (Figure 2 and Figure 3), higher deciles of the predicted outcomes are associated with higher observed outcomes, which show a good calibration of the model. The plot indicates good discrimination of the model and good predictive ability.

Overall, these diagnostic results demonstrate the risk-adjustment model adequately controls for differences in patient characteristics (case mix).

Table 8. Model Performance of Risk-Adjusted Model of SCB Improvement Following THA/TKA

Model Performance Statistic	Development Dataset	Validation Dataset
C-statistic	0.61	0.61
Calibration (γ0, γ1)	(0, 1)	(0.02, 0.97)
Discrimination — Predictive ability (lowest decile %, highest decile %)	(48.67%, 80.03%)	(52.44%, 81.14%)

Table 9. Risk-Adjustment Model Performance Residuals Lack of Fit

Residuals Lack of Fit (Pearson Residual %)	Development Dataset	Validation Dataset
<-2	104 (0.89%)	58 (0.75%)
[-2,0)	3739 (32.09%)	2483 (31.93%)
[0,2)	7810 (67.02%)	5235 (67.32%)
[2 +)	0 (0.00%)	0 (0.00%)

Figure 2. Calibration Deciles for the Development Dataset

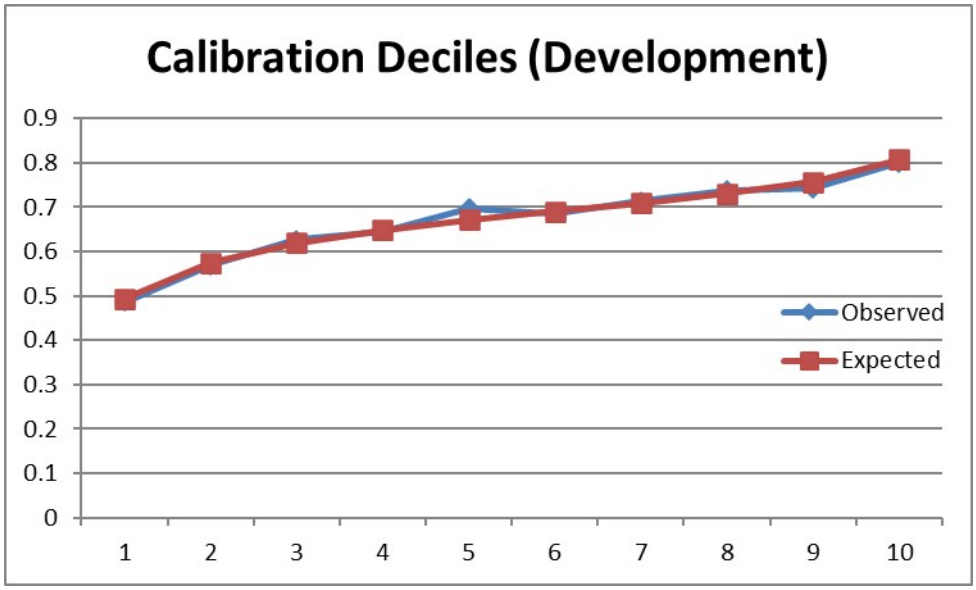
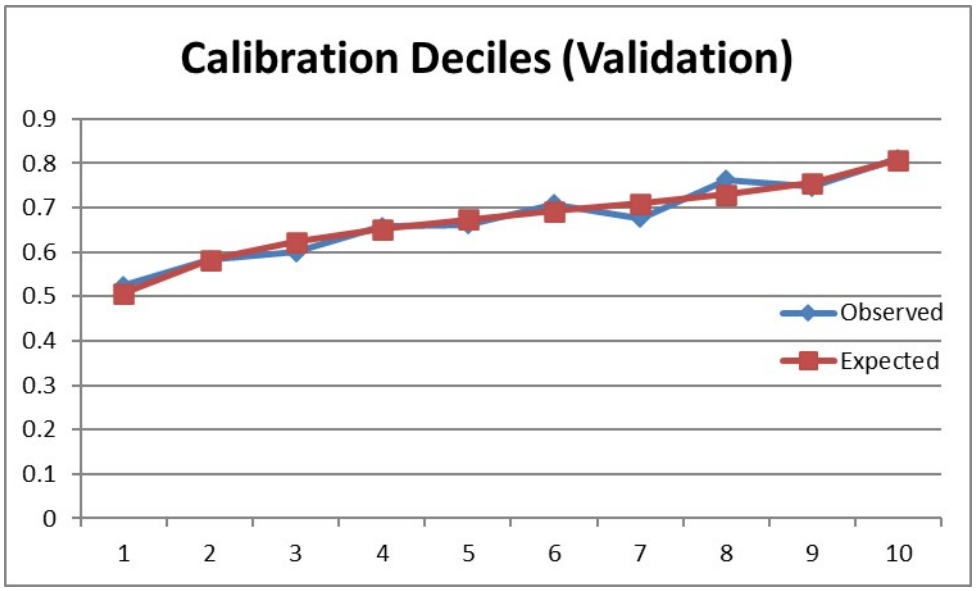


Figure 3. Calibration Deciles for the Validation Dataset



In additional analyses, we compared the C-statistic in the clinician- and clinician group-level measure to the C-statistic in the hospital-level THA/TKA PRO-PM. These analyses revealed that the risk model for the clinician and clinician group measure resulted in a C-statistic that was lower than that of the risk model for the hospital-level measure and led to examination of risk model performance per each of the two years of procedure data for the clinician and clinician group measure. Per the results in Table 10, the C-statistic using the same data as that used to calculate the hospital measure (only the first full year of PRO data) was 0.68, the same as the risk model C-statistic for the hospital measure. However, the risk model for the clinician and clinician group measure using only the second full year of data revealed a

lower C-statistic (.059), revealing a difference in risk model prediction between risk models using the first and second years' THA/TKA procedures with PRO data. Since the risk model is performed at the patient-level, we do not anticipate the clinician attribution would impact the results. Investigation of the patient characteristics between those undergoing a THA/TKA in the first year (July 1, 2016 – June 30, 2017) and those undergoing a THA/TKA in the second year (July 1, 2017 – June 30, 2018) did reveal a strong association between improvement and health literacy in the first year of data that does not persist in the second year of data. We recommend continued assessment of the risk model in a larger dataset in the future.

Table 10. Risk Model Performance by Procedure Year: C-Statistics Comparison

Model Performance Statistic	C-statistics	Number of Patients	Mean Proportion of Patients with SCB
Clinician- and clinician group-level measure (July 1, 2016 – June 30, 2018, two full years of PRO data)	0.61	19,429	67.14%
Clinician- and clinician group-level measure (July 1, 2016 – June 30, 2017, 1st full year of PRO data)	0.68	9,905	64.37%
Clinician- and clinician group-level measure (July 1, 2017 – June 30, 2018, 2nd full year of PRO data)	0.59	9,524	70.02%

3.3.1 Social Risk Factor Assessment

Bivariate and multivariate analyses conducted in the Development Dataset showed no statistically significant association between dual eligibility, AHRQ SES index lowest quartile, or non-white race with SCB improvement at the bivariate level (Table 11) or when entered into the risk model (Table 12). Tables 13 and 14 provide the mean and range of clinician- and clinician group-specific RSIRs with no social risk factors included in the risk model, with dual eligibility, AHRQ SES index lowest quartile, and non-white race patients individually included in the risk model, and with all three social risk factors included in the model together. Correlation coefficients between RSIRs calculated without social risk factors compared to RSIRs calculated with each of the social risk factors entered individually as well as all three social risk factors included in the model together indicates near-perfect correlation in our data.

Table 11. Bivariate Associations of Social Risk Factors and Race with SCB Improvement: Development Dataset (Patient N = 11,653)

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

Table 12. Adjusted ORs for Social Risk Factors and Race Individually Evaluated in the Risk Model for SCB Improvement: Development Dataset (Patients 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

* C-statistic for the risk model for SCB improvement in the Development Dataset without any of the three social risk factors = 0.61

Table 13. 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 Procedures with 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
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

Table 14. 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 Procedures with 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

Analysis of proportion of dual eligible, AHRQ SES index lowest quartile, and non-white race patients by clinician and clinician group RSIRs is provided in Figures 4 through 9. The results indicate that clinicians and clinician groups with the lowest proportion of dual eligible, AHRQ SES index lowest quartile, and non-white race patients and those clinicians and clinician groups with the highest proportion of these patients have similar RSIR distributions. These data do not provide evidence of significant differences in RSIRs across clinician or clinician groups with different proportions of patients with dual eligibility, AHRQ SES index lowest quartile, and non-white race.

Figure 4. MIPS THA/TKA PRO-PM RSIRs by Proportion of Dual Eligible Patients (Clinicians)

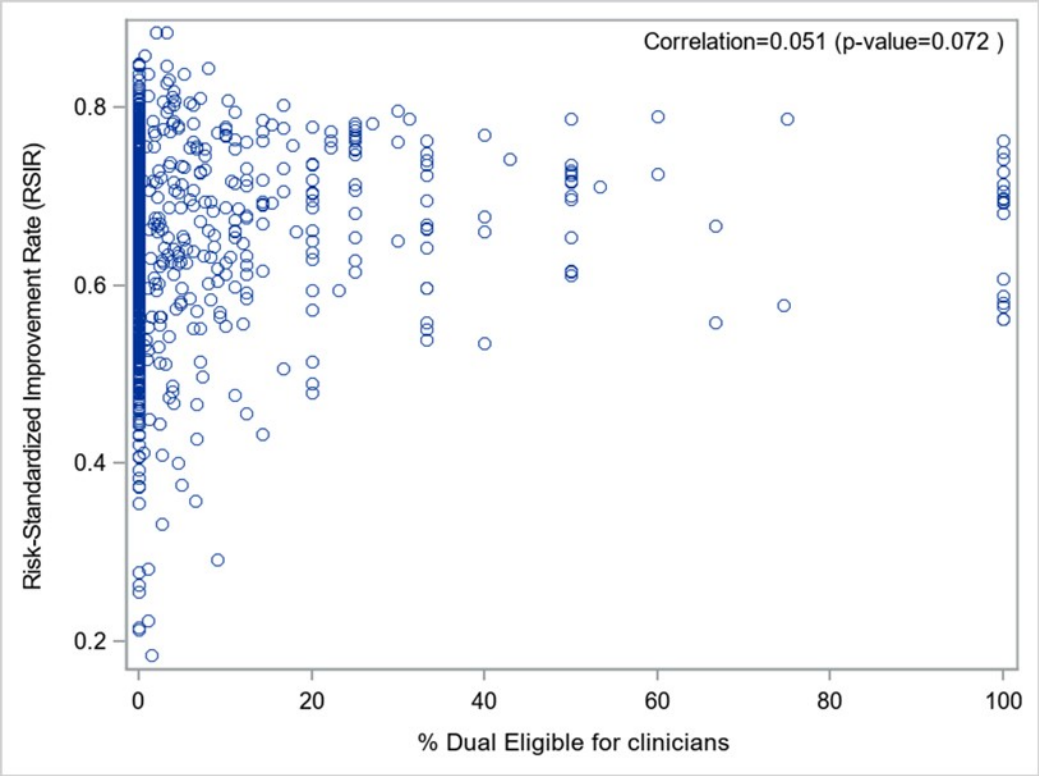


Figure 5. MIPS THA/TKA PRO-PM RSIRs by Proportion of Dual Eligible Patients (Clinician Groups)

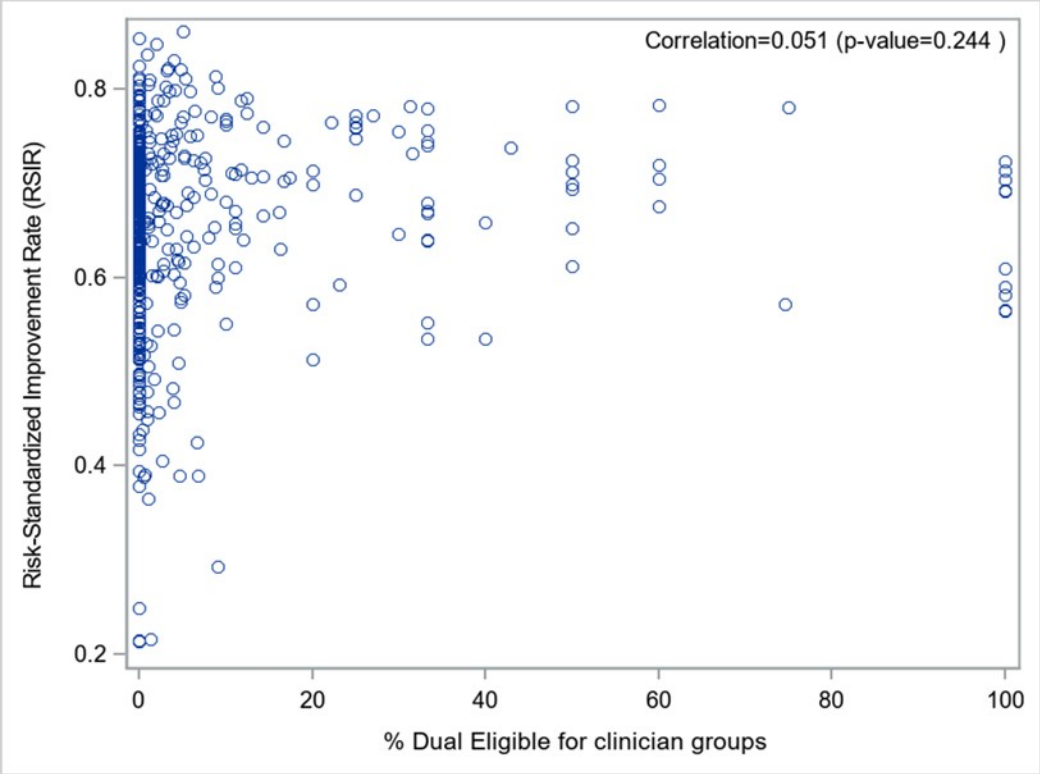


Figure 6. MIPS THA/TKA PRO-PM RSIRs by Proportion of AHRQ SES Index Lowest Quartile Patients (Clinicians)

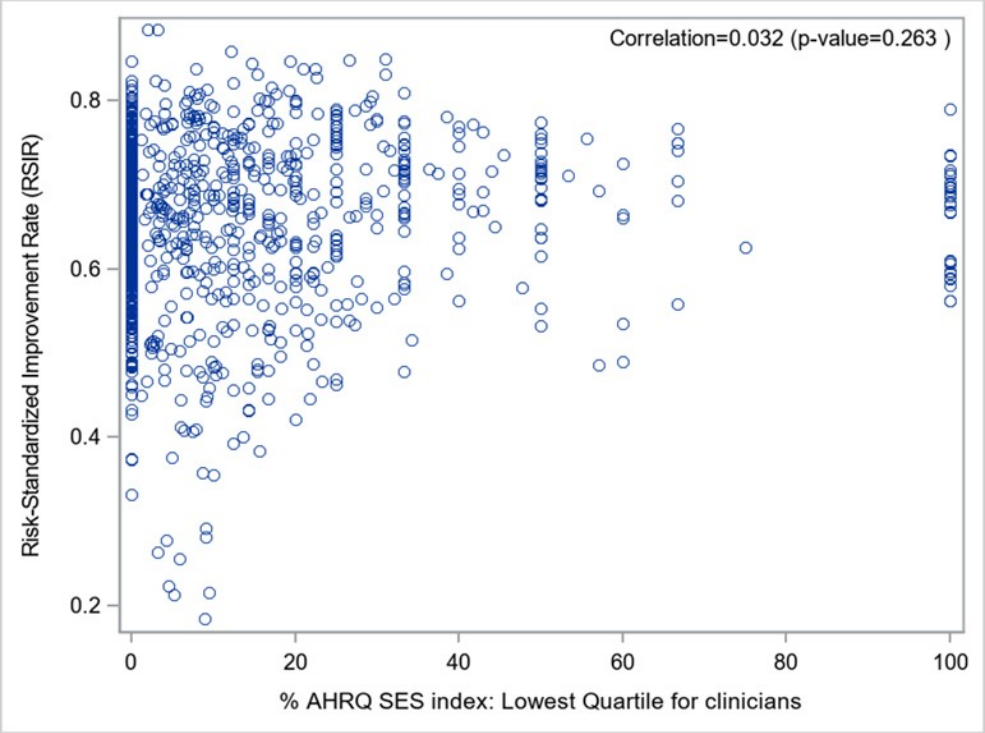


Figure 7. MIPS THA/TKA PRO-PM RSIRs by Proportion of AHRQ SES Index Lowest Quartile Patients (Clinician Groups)

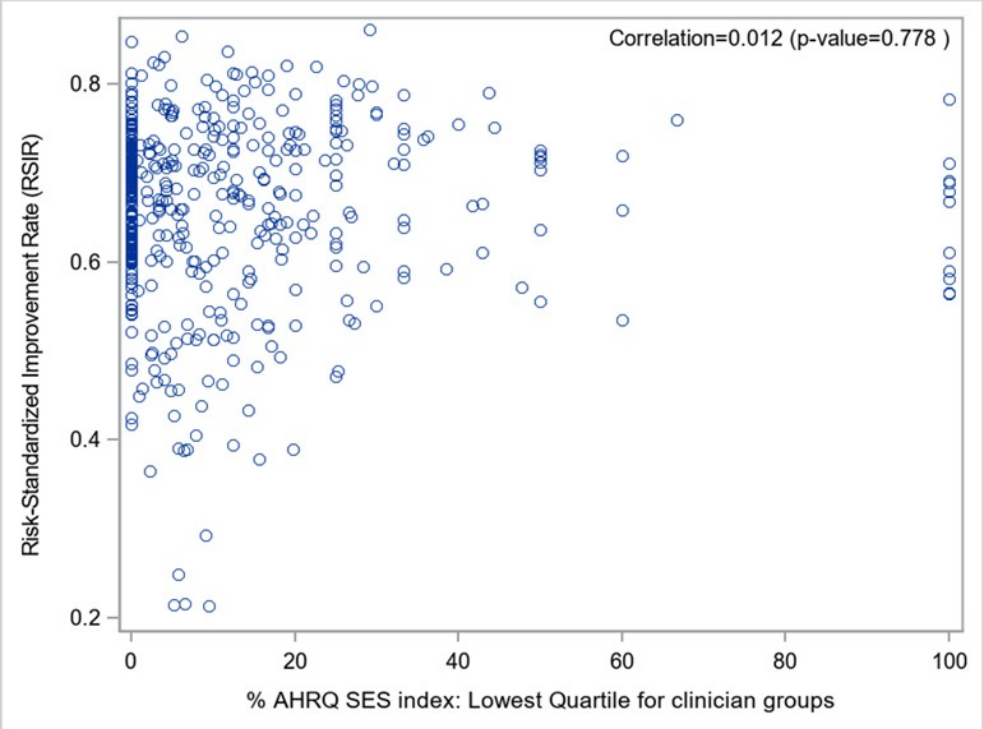


Figure 8. MIPS THA/TKA PRO-PM RSIRs by Proportion of Non-white Race Patients (Clinicians)

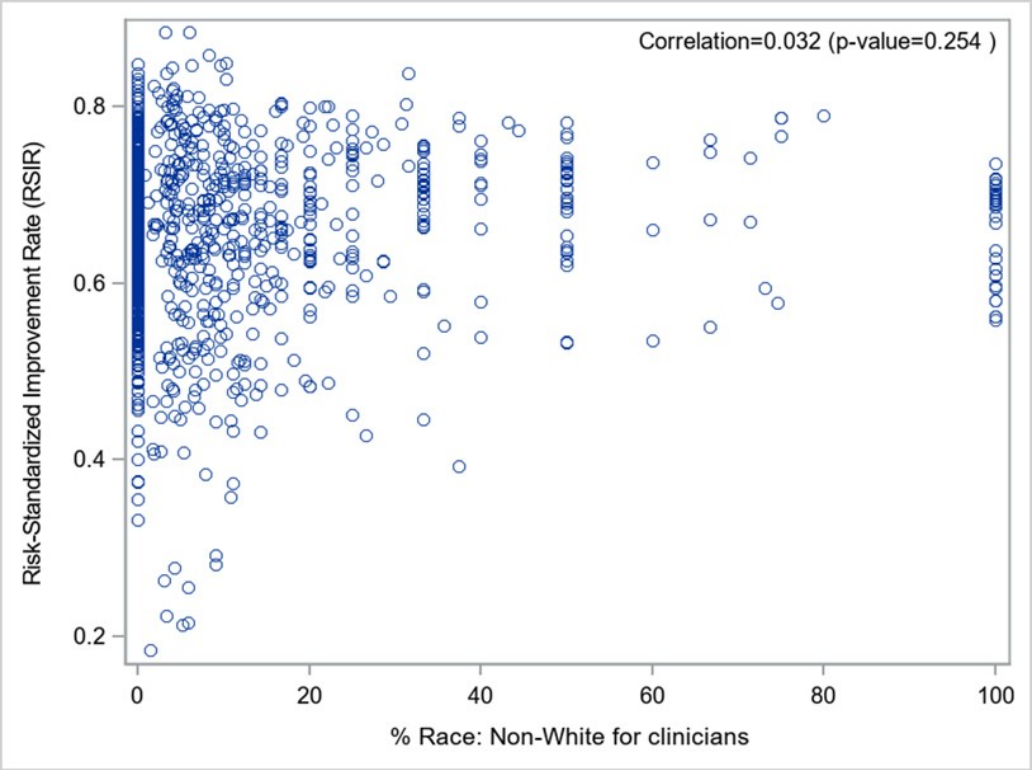
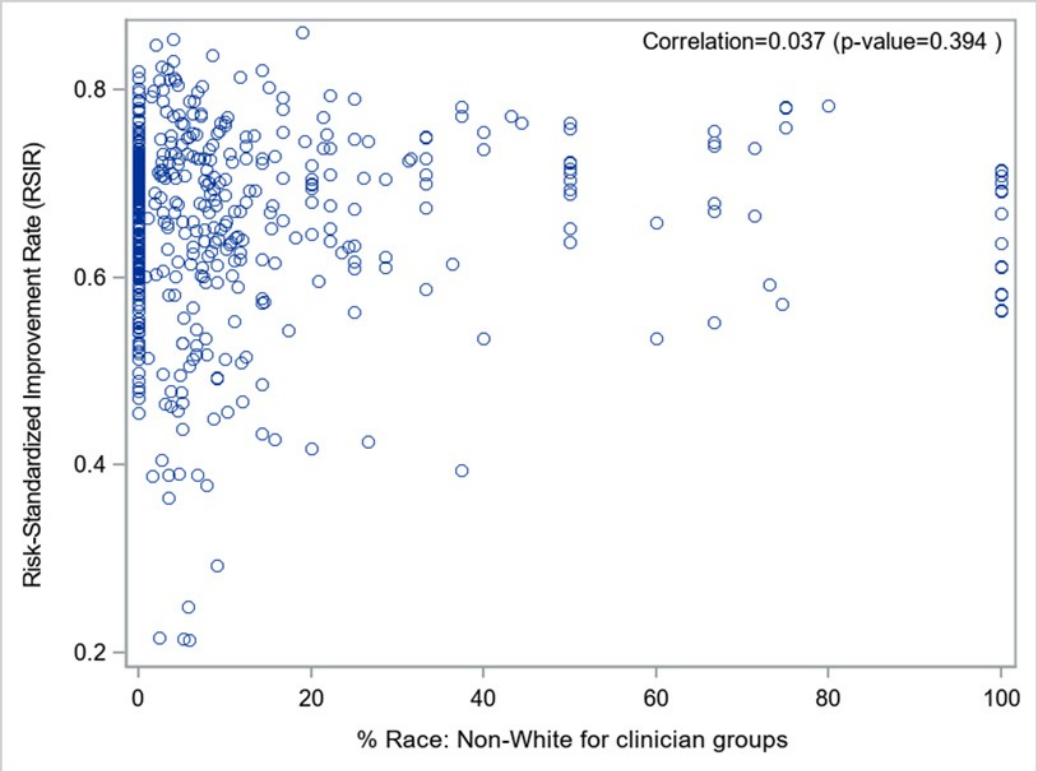


Figure 9. MIPS THA/TKA PRO-PM RSIRs by Proportion of Non-white Race Patients (Clinician Groups)



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. In our dataset, dual eligibility, AHRQ SES index lowest quartile, and non-white race were not statistically significantly associated with the outcome, and inclusion of these variables in the risk model did not appear to impact RSIRs. The lack of association and effect of these factors may be due to lower case selection in these groups for these elective primary procedures.

3.4 Measure Results

Table 15 provides the mean and distribution of clinician and clinician group 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%). The variation in RSIRs suggests meaningful differences exist 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.

Table 15. Mean and Distribution of RSIRs for Risk Model of SCB Improvement Following Elective Primary THA/TKA (Clinicians and Clinician Groups with ≥25 THA/TKA Procedures with PRO Data)

Summary Statistics	Clinician-level RSIRs (Combined Dataset)	Clinician Group-level RSIRs (Combined Dataset)
N	232 (Clinicians)	170
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%

3.5 Reliability Results

3.5.1 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 MIPS THA/TKA PRO-PM is based.

HOOS, JR Reliability

Internal Consistency: The developers of the HOOS, JR[2] 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 the six HOOS, JR items existed in a single dimension.[2]

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.[60-63] 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[3] assessed internal consistency reliability of using the PSI. Internal consistency of the KOOS, JR on the PSI were 0.84 in the HSS cohort and 0.85 in the 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.[3]

Test-retest Reliability: Test-retest reliability was not tested by developers of the KOOS, JR as it had already been tested in the KOOS.[64] 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).

The reliability results from the literature demonstrate 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[2] and 0.84 – 0.85 for the KOOS, JR.[3] These values are well above 0.7 affirm the instruments' ability to differentiate patients with varying levels of pain and functioning, which 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.

3.5.2 Measure Score Reliability Results

Table 16 shows the signal-to-noise reliability for clinicians and clinician groups with at least 5, 10, and 25 THA/TKA procedures with complete PRO data. At the thresholds of at least 5 and 10 procedures with

complete PRO data, the median reliability scores were all at or above 0.7 indicating acceptable reliability. At the threshold of at least 25 procedures with complete PRO data, the median reliability scores were 0.87 (range 0.79– 0.97) for clinicians and 0.92 (range: 0.79 – 0.99) for clinician groups, indicating excellent reliability.

Table 16. Signal-to-Noise-Reliability, Clinicians and Clinician Groups

Attribution and Volume	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 ≥20 THA/TKA Procedures with Complete PRO Data	286	0.85	0.85 (0.06)	0.76	0.97	0.80	0.90	0.10
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 ≥20 THA/TKA Procedures with Complete PRO Data	193	0.90	0.89 (0.07)	0.75	0.99	0.83	0.95	0.12

Attribution and Volume	N	Median	Mean (SD)	Min	Max	Interquartile Range		
						Q1	Q3	Range
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

The signal-to-noise reliability of 0.87 and 0.92 for clinicians and clinician groups, respectively indicates excellent reliability. Our interpretation of these results is based on standards established by Landis and Koch[67]:

- <0 = Less than chance agreement
- 0 – 0.2 = Slight agreement
- 0.21 – 0.39 = Fair agreement
- 0.4 – 0.59 = Moderate agreement
- 0.6 – 0.79 = Substantial agreement
- 0.8 – 0.99 = Almost Perfect agreement
- 1 = Perfect agreement

3.6 Validity Results

3.6.1 Data Element Validity Results

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

HOOS, JR Validity

Responsiveness: Standardized response means for the HOOS, JR relative to other PROMs measuring postoperative 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-TJR registry data.[2]

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-TJR 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-TJR 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-TJR registry data. Between HOOS, JR and WOMAC Functioning domain, correlations 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-TJR registry data. Between the HOOS, JR and WOMAC Stiffness domain, correlations 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-TJR registry data.[2]

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.[2]

KOOS, JR Validity

Responsiveness: Standardized response means for the KOOS, JR relative to other PROMs measuring postoperative 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.[3]

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-TJR 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-TJR registry data. Correlations with the KOOS 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-TJR 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-TJR registry data. Between KOOS, JR and WOMAC Function, correlations 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-TJR registry data. Between KOOS, JR and WOMAC Stiffness, correlations 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-TJR registry data.[3]

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%.[3]

The validity results from the literature demonstrate the HOOS, JR and the KOOS, JR PROM instruments are valid and meaningful measures for assessing PROs 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."[65] 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 KOOS 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.[2, 3]

3.6.3 Face Validity Results

Among TEP members who provided responses, 100% of members agreed with the first statement ("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") and 88.2% agreed with the second statement ("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"). Among Patient Working Group members who provided responses, 100% of members agreed with both statements.

For the first statement, among the 17 TEP members who provided responses, 7 responded "Strongly Agree", 6 responded "Moderately Agree", and 4 responded "Somewhat Agree". Among the 4 Patient

Working Group members who provided responses, 2 responded “Strongly Agree” and 2 responded “Moderately Agree” to the first statement.

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

In summary, 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. In addition, Patient Working Group members supported the measure concept and agreed that the measure results would provide meaningful information to them or other patients.

3.7 Missing Data

Accounting for missing data is considered in the response bias adjustment described in Section 2.7.1.

3.8 Response Rates and Response Bias Adjustment

PRO submission rates for all clinicians and clinician groups and for clinicians and clinician groups with 25 or more THA/TKA procedures with PRO data are presented in Table 17.

Table 17. PRO Submission Rates

Summary Statistics	PRO Submission Rates (All Clinicians)	PRO Submission Rates (Clinicians with ≥25 THA/TKA Procedures with PRO Data)	PRO Submission Rates (All Clinician Groups)	PRO Submission Rates (Clinician Groups with ≥25 THA/TKA Procedures with PRO Data)
N	1254	232	526	170
Mean (SD)	32.23% (24.55)	42.09% (16.98)	31.85% (24.20)	36.65% (18.38)
Percentile	-	-	-	-
100% Max	100.00%	89.47%	100.00%	84.48%
99%	100.00%	85.44%	100.00%	84.44%
95%	77.78%	70.73%	77.78%	68.18%
90%	66.67%	62.26%	65.38%	60.25%
75% Q3	50.00%	54.30%	48.91%	50.54%
50% Median	27.07%	41.82%	27.40%	36.31%
25% Q1	11.43%	27.84%	11.11%	21.33%
10%	4.31%	19.61%	4.31%	13.70%

Summary Statistics	PRO Submission Rates (All Clinicians)	PRO Submission Rates (Clinicians with ≥25 THA/TKA Procedures with PRO Data)	PRO Submission Rates (All Clinician Groups)	PRO Submission Rates (Clinician Groups with ≥25 THA/TKA Procedures with PRO Data)
5%	2.50%	16.83%	2.64%	8.77%
1%	0.96%	12.25%	0.96%	4.35%
0% Min	0.31%	10.98%	0.31%	2.89%

Propensity scores calculated using a multinomial logistic regression where the outcome was 1) complete PRO submission, 2) incomplete PRO submission, and 3) no response were used to calculate stabilized IPW for each of these three groups. Table 18 provides the distribution of these stabilized weights applied to procedures with complete PRO data submission in the risk adjustment model.

Table 18. Distribution of Stabilized Weights Applied to Procedures with Complete PRO Submission in the Risk Adjustment Model (Responders)

Summary Statistics	Stabilized Weights
Mean (SD)	0.9996 (0.2539)
Percentile	-
100% Max	4.95759
99%	1.65032
95%	1.2759
90%	1.14406
75% (Q3)	1.03151
50% (Median)	0.95572
25% (Q1)	0.89917
10%	0.85254
5%	0.82368
1%	0.76378
0% Min	0.56802

We assessed the non-response bias by the Pearson correlation between the residuals of the hierarchical outcome model with only clinical risk factors and the probability of response. For clinicians, this correlation is -0.00784 (p-value 0.27). For clinician groups, this correlation is it is -0.00709 (p-value 0.32).

This indicates that there is not an association between the residuals of the improvement outcome and the probability of response based on our models.

The comparison of clinician-level and clinician group-level RSIRs for a risk-adjusted model of SCB improvement with stabilized IPW and without stabilized IPW (Table 19 and Table 20) suggests that the results are not sensitive to our weighting adjustment. Due to the high proportion of non-responders, however, we considered it important to account for the differences in characteristics of responders and non-responders found in the literature and empirically in our data. We expect non-response bias will be a factor for the MIPS THA/TKA PRO-PM due to associations with non-response including SES and health status. We, therefore, retained response bias adjustment for the measure results.

Table 19. Mean and Distribution of Clinician RSIRs for Risk-Adjusted Model of SCB Improvement with and without Stabilized Inverse Probability Weighting for Potential Non-Response Bias (Combined Dataset, Clinicians with ≥25 THA/TKA Procedures with PRO Data)

Summary Statistics	RSIRs (No Weighting)	RSIRs (Weighted for Non-Response)
N (Clinician)	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%
5%	41.40%	40.98%
1%	22.31%	22.33%
0% Min	18.36%	18.45%

Table 20. Mean and Distribution of Clinician Group RSIRs for Risk-Adjusted Model of SCB Improvement with and without Stabilized Inverse Probability Weighting for Potential Non-Response Bias (Combined Dataset, Clinician Groups with ≥25 THA/TKA Procedures with PRO Data)

Summary Statistics	RSIRs (No Weighting)	RSIRs (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%

4. Report Conclusions

In summary, we report the specifications and testing results for a clinician- and clinician group-level PRO-PM for patients undergoing elective primary THA/TKA. The measure has been fully tested and is ready for submission to NQF for measure endorsement. The measure specifications are aligned with the hospital-level THA/TKA PRO-PM and call for the use of procedure-specific PROMs (HOOS, JR or KOOS, JR) and 19 risk variables including patient-reported health literacy, back pain, lower extremity joint pain, and mental health (using the PROMIS-Global or VR-12). The measure cohort includes patients undergoing elective primary THA/TKA (not including patients with hip and/or pelvic fractures and/or revision THAs/TKAs). The measure outcome assesses achievement of SCB improvement between preoperative and postoperative assessment of PROM scores (for THA patients: meeting or exceeding the SCB threshold of 22 points on the HOOS, JR and for TKA patients: meeting or exceeding the SCB threshold of 20 points on the KOOS, JR.). The preoperative data collection timeframe will be 90 to zero days before the procedure and the postoperative data collection timeframe will be 300 to 425 days following the procedure. The measure addresses potential response bias using stabilized inverse probability weighting.

Finally, stakeholders over time shared feedback regarding measurement of THA and TKA procedures performed in Outpatient (OP) and Ambulatory Surgical Center (ASC) settings given an increasing number of procedures are being performed in these settings due to recent CMS regulations.[68] We also heard support for measurement of THA/TKA procedures across settings during discussion of the hospital-level THA/TKA PRO-PM at the Rural Measures Application Partnership Meeting in January 2021. Expanding the cohort to procedures performed in the OP and ASC settings may increase the number of eligible patients and providers included in the measure results and decrease a possible unintended consequence of diverting procedures to certain settings to avoid measurement. CMS plans to consider this measure for MIPS which is agnostic to settings; therefore, implementing a measure which includes procedures across settings is feasible. Our testing dataset did not include procedures in the OP and ASC settings. CMS will evaluate the impact of expanding the applicable settings during implementation planning.

References

1. Balestracci, K., et al. *Patient-Reported Outcomes (PROs) Following Elective Primary Total Hip and/or Total Knee Arthroplasty: Hospital-Level Performance Measure* 2020; Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Measure-Methodology.
2. Lyman, S., et al., *Validation of the HOOS, JR: A Short-form Hip Replacement Survey.* Clin Orthop Relat Res, 2016. **474**(6): p. 1472-82.
3. Lyman, S., et al., *Validation of the KOOS, JR: A Short-form Knee Arthroplasty Outcomes Survey.* Clin Orthop Relat Res, 2016. **474**(6): p. 1461-71.
4. National Quality Forum. *Patient Reported Outcomes (PROs) in Performance Measurement* 2013; Available from: http://www.qualityforum.org/Projects/n-r/Patient-Reported_Outcomes/Patient-Reported_Outcomes.aspx.
5. Centers for Medicare and Medicaid Services, *Medicare Program; Merit-Based Incentive Payment System (MIPS) and Alternative Payment Model (APM) Incentive Under the Physician Fee Schedule, and Criteria for Physician-Focused Payment Models. Final rule with comment period.* Fed Regist, 2016. **81**(214): p. 77008-831.
6. Zhang, Y. and J.M. Jordan, *Epidemiology of osteoarthritis.* Clin Geriatr Med, 2010. **26**(3): p. 355-69.
7. Centers for Disease Control and Prevention. *Osteoarthritis (OA)*. 2020; Available from: https://www.cdc.gov/arthritis/basics/osteoarthritis.htm.
8. Guccione, A.A., et al., *The effects of specific medical conditions on the functional limitations of elders in the Framingham Study.* Am J Public Health, 1994. **84**(3): p. 351-8.
9. Michaud, C.M., et al., *The burden of disease and injury in the United States 1996.* Popul Health Metr, 2006. **4**: p. 11.
10. Levit, K., et al. *HCUP Facts and Figures, 2006: Statistics on Hospital-based Care in the United States*. 2008; Available from: https://www.hcup-us.ahrq.gov/reports/factsandfigures/facts_figures_2006.jsp.
11. Healthcare Cost and Utilization Project. *HCUP Fast Stats - Most Common Diagnoses for Inpatient Stays* 2021; Available from: https://www.hcup-us.ahrq.gov/faststats/NationalDiagnosesServlet?year1=2018&characteristic1=0&included1=1&year2=2017&characteristic2=0&included2=1&expansionInfoState=hide&dataTablesState=hide&definitionsState=hide&exportState=hide.
12. Liang, L., B. Moore, and A. Soni, *National Inpatient Hospital Costs: The Most Expensive Conditions by Payer, 2017. HCUP Statistical Brief #261.* 2020.
13. Lopez, C.D., et al., *Hospital and Surgeon Medicare Reimbursement Trends for Total Joint Arthroplasty.* Arthroplast Today, 2020. **6**(3): p. 437-444.
14. Rissanen, P., et al., *Health and quality of life before and after hip or knee arthroplasty.* The Journal of Arthroplasty, 1995. **10**(2): p. 169-175.
15. Wiklund, I. and B. Romanus, *A comparison of quality of life before and after arthroplasty in patients who had arthrosis of the hip joint.* J Bone Joint Surg Am, 1991. **73**(5): p. 765-9.
16. Laupacis, A., et al., *The effect of elective total hip replacement on health-related quality of life.* J Bone Joint Surg Am, 1993. **75**(11): p. 1619-26.
17. Ritter, M.A., et al., *Comparative outcomes of total joint arthroplasty.* The Journal of Arthroplasty, 1995. **10**(6): p. 737-741.

18. Liebs, T.R., 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**(2): p. 239-43.
19. DeBuhr, J., et al., *2021 Procedure-Specific Readmission Measures Updates and Specifications Report: Isolated Coronary Artery Bypass Graft (CABG) Surgery - Version 8.0, Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) - Version 10.0* 2021.
20. National Joint Registry. *National Joint Registry for England and Wales 9th Annual Report 2012*. 2012; Available from: www.njrcentre.org.uk.
21. Ghomrawi, H.M.K., et al., *How Often are Patient and Surgeon Recovery Expectations for Total Joint Arthroplasty Aligned? Results of a Pilot Study.* HSS Journal ®, 2011. **7**(3): p. 229-234.
22. Harris, I.A., et al., *Discordance Between Patient and Surgeon Satisfaction After Total Joint Arthroplasty.* The Journal of Arthroplasty, 2013. **28**(5): p. 722-727.
23. Jourdan, C., et al., *Comparison of Patient and Surgeon Expectations of Total Hip Arhtroplasty.* PLoS ONE, 2012. **7**(1): p. e30195.
24. Suda, A.J., et al., *Are patients' expectations of hip and knee arthroplasty fulfilled? A prospective study of 130 patients.* Orthopedics, 2010. **33**(2): p. 76-80.
25. American Academy of Orthopaedic Surgeons, *Preventing Venous Thromboembolic Disease in Patients Undergoing Elective Hip and Knee Arthroplasty: Evidence-Based Guideline and Evidence Report*. 2011.
26. Anderson, F.A., et al., *Prevention of Venous Thromboembolism After Hip or Knee Arthroplasty: Findings From a 2008 Survey of US Orthopedic Surgeons.* The Journal of Arthroplasty, 2012. **27**(5): p. 659-666.e5.
27. Roos, E.M., *Effectiveness and practice variation of rehabilitation after joint replacement.* Curr Opin Rheumatol, 2003. **15**(2): p. 160-2.
28. National Health System. *HESonline Hospital Episode Statistics: Proms Data*. 2012; Available from: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.
29. Neuburger, J., et al., *Using patient-reported outcomes (PROs) to compare the providers of surgery: does the choice of measure matter?* Med Care, 2013. **51**(6): p. 517-23.
30. DeBuhr, J., et al., *2021 Procedure-Specific Complication Measure Updates and Specifications Report: Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) (Version 10.0)* 2021.
31. Herrin, J., et al. *Measure Methdology Report: Risk-standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) for the Merit-based Incentive Payment System (MIPS)*. 2020; Available from: https://qpp.cms.gov/mips/explore-measures?tab=qualityMeasures&py=2021.
32. Pincus, D., et al., *Association Between Surgical Approach and Major Surgical Complications in Patients Undergoing Total Hip Arthroplasty.* Jama, 2020. **323**(11): p. 1070-1076.
33. Lau, R.L., 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): p. 250.
34. Malik, A.T., et al., *Does Surgeon Volume Affect Outcomes Following Primary Total Hip Arthroplasty? A Systematic Review.* The Journal of Arthroplasty, 2018. **33**(10): p. 3329-3342.
35. Levaillant, M., et al., *Assessing the hospital volume-outcome relationship in surgery: a scoping review protocol.* BMJ Open, 2020. **10**(10): p. e038201.
36. Black, N., M. Varaganum, and A. Hutchings, *Relationship between patient reported experience (PREMs) and patient reported outcomes (PROMs) in elective surgery.* BMJ Quality & Safety, 2014. **23**(7): p. 534.

37. Centers for Medicare and Medicaid Services. *2020 Merit-based Incentive Payment System (MIPS) Quality Measures List* 2019; Available from: https://qpp.cms.gov/mips/explore-measures?tab=qualityMeasures&py=2020.
38. Zuckerman, D.M., P. Brown, and S.E. Nissen, *Medical device recalls and the FDA approval process.* Arch Intern Med, 2011. **171**(11): p. 1006-11.
39. Centers for Medicare and Medicaid Services. *Meaningful Measures Hub* 2021; Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityInitiativesGenInfo/MMF/General-info-Sub-Page.
40. Centers for Medicare and Medicaid Services, *Blueprint for the CMS Measures Management System - Version 16.0*. 2020.
41. National Quality Forum, *National Voluntary Consensus Standards for Patient Outcomes 2009: A Consensus Report* 2011.
42. National Quality Forum, *Patient-Reported Outcomes: Best Practices on Selection and Data Collection. Final Technical Report*. 2020.
43. Krumholz, H.M., et al., *Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation.* Circulation, 2006. **113**(3): p. 456-62.
44. Suter, L., et al., *Environmental Scan/Literature Review: Total Hip and Total Knee Arthroplasty Patient-Reported Outcome Measure*. 2012.
45. Lyman, S., et al., *What Are the Minimal and Substantial Improvements in the HOOS and KOOS and JR Versions After Total Joint Replacement?* Clin Orthop Relat Res, 2018. **476**(12): p. 2432-2441.
46. Pope, G., et al., *Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment: Final Report*. 2000.
47. Pope, G., et al., *Evaluation of the CMS-HCC Risk Adjustment Model: Final Report*. 2011.
48. Fairbank, J.C. and P.B. Pynsent, *The Oswestry Disability Index.* Spine (Phila Pa 1976), 2000. **25**(22): p. 2940-52; discussion 2952.
49. 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.* J Bone Joint Surg Am, 2013. **95**(20): p. 1833-7.
50. Irgit, K. and C.L. Nelson, *Defining Racial and Ethnic Disparities in THA and TKA.* Clinical Orthopaedics and Related Research®, 2011. **469**(7): p. 1817-1823.
51. Kerman, H.M., et al., *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 Res (Hoboken), 2018. **70**(9): p. 1326-1334.
52. Hutchings, A., et al., *Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England.* Health Qual Life Outcomes, 2012. **10**: p. 34.
53. de Rooij, B.H., et al., *Cancer survivors not participating in observational patient-reported outcome studies have a lower survival compared to participants: the population-based PROFILES registry.* Qual Life Res, 2018. **27**(12): p. 3313-3324.
54. Schamber, E.M., et al., *Barriers to completion of Patient Reported Outcome Measures.* J Arthroplasty, 2013. **28**(9): p. 1449-53.
55. Garrido, M.M., *Covariate Adjustment and Propensity Score.* Jama, 2016. **315**(14): p. 1521-2.
56. Thoemmes, F. and A.D. Ong, *A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models* Emerging Adulthood 2015. **4**(1): p. 40-59.

57. Patel, J., et al., *Predictors of Low Patient-Reported Outcomes Response Rates in the California Joint Replacement Registry.* J Arthroplasty, 2015. **30**(12): p. 2071-5.
58. Harrell, F.E., Jr. and Y.C. Shih, *Using full probability models to compute probabilities of actual interest to decision makers.* Int J Technol Assess Health Care, 2001. **17**(1): p. 17-26.
59. Merlo, J., et al., *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, 2006. **60**(4): p. 290-7.
60. de Groot, I.B., et al., *Validation of the Dutch version of the Hip disability and Osteoarthritis Outcome Score.* Osteoarthritis Cartilage, 2007. **15**(1): p. 104-9.
61. Klässbo, M., E. Larsson, and E. Mannevik, *Hip disability and osteoarthritis outcome score. An extension of the Western Ontario and McMaster Universities Osteoarthritis Index.* Scand J Rheumatol, 2003. **32**(1): p. 46-51.
62. Nilsdotter, A. and A. Bremander, *Measures of hip function and symptoms: Harris Hip Score (HHS), Hip Disability and Osteoarthritis Outcome Score (HOOS), Oxford Hip Score (OHS), Lequesne Index of Severity for Osteoarthritis of the Hip (LISOH), and American Academy of Orthopedic Surgeons (AAOS) Hip and Knee Questionnaire.* Arthritis Care Res (Hoboken), 2011. **63 Suppl 11**: p. S200-7.
63. Ornetti, P., et al., *Cross-cultural adaptation and validation of the French version of the Hip disability and Osteoarthritis Outcome Score (HOOS) in hip osteoarthritis patients.* Osteoarthritis Cartilage, 2010. **18**(4): p. 522-9.
64. Roos, E.M., et al., *Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure.* J Orthop Sports Phys Ther, 1998. **28**(2): p. 88-96.
65. Norman, G. and D. Streiner, *Health Measurement Scales: A Practical Guide to Their Development and Use* 3rd ed. 2003: Oxford University Press.
66. Wechsler, S., *Statistics at Square One*. 9th ed. 1997, London, UK: BMJ Publishing Group 140.
67. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
68. Centers for Medicare and Medicaid Services, *Medicare Program: Hospital Outpatient Prospective Payment and Ambulatory Surgical Center Payment Systems and Quality Reporting Programs; Price Transparency of Hospital Standard Charges; Radiation Oncology Model; Request for Information on Rural Emergency Hospitals.* Fed Regist, 2021. **86**(147): p. 42251-42252, 42277.

Appendices

A. Acknowledgement Details

CORE Expert Clinical Consultant:

Kevin Bozic, MD, MBA
Dell Medical School at the University of Texas, Austin (Professor and Chair of the Department of Surgery and Perioperative Care); Austin, TX

Technical Expert Panel Members (2020 – 2021): of note, TEP members identified for years prior to 2020 participated on the hospital-level THA/TKA PRO-PM TEP

David C. Ayers, MD
University of Massachusetts (UMass) Medical School (Professor and Chair of Orthopedics and Physical Rehabilitation); Worcester, MA
(TEP member 2013 – 2021)

Thomas C. Barber, MD
Memorial Sloan Kettering Hospital (Deputy Physician in Chief); New York, NY
(TEP member 2013 – 2021)

Phyllis Bass
Patient Expert; Cypress, TX
(TEP member 2020 – 2021)

Vinod Dasa, MD
Louisiana State University Health Sciences Center (Professor of Clinical Orthopedics, Director of Research); New Orleans, LA
(TEP member 2013 – 2021)

Rachel DuPré Brodie
Purchaser Business Group on Health (Senior Director of Measurement & Accountability); San Francisco, CA
(TEP member 2018 – 2021)

Cheryl Fahlman, PhD, MBA, BSP
CAF Consulting Solutions (President); Gaithersburg, MD
(TEP member 2013 – 2021)

William G. Hamilton, MD
Inova Mount Vernon Hospital and American Association of Hip and Knee Surgeons (Chair of Orthopedic Surgery; FOCAL Chair); Alexandria, VA
(TEP member 2018 – 2021)

Cynthia S. Jacelon, PhD, RN-BC, CRRN, FGSA, FAAN
University of Massachusetts Amherst School of Nursing (Professor and Executive Associate Dean); Greenfield, MA
(TEP member 2013 – 2021)

Patient B
Patient Expert; Chicago, IL
(TEP member 2019 – 2021)

Patient A
Patient Expert; Sherman Oaks, CA
(TEP member 2019 – 2021)

Craig T. Miller, PT
Rivetus Rehabilitation and American Physical Therapy Association (Director of Home Care Therapy and Senior PT); Macomb, MI
(TEP member 2021)

Michael H. Perskin, MD
The American Geriatrics Society; New York University School of Medicine (Clinical Professor of Medicine); New York, NY
(TEP member 2013 – 2021)

Nan Rothrock, PhD
Feinberg School of Medicine of Northwestern University (Associate Professor of Medical Social Sciences); Chicago, IL
(TEP member 2018 – 2021)

Jonathan L. Schaffer, MD, MBA, FACS, FHIMSS, FABOS
The Cleveland Clinic (Staff and Program Director); Cleveland, OH
(TEP member 2013 – 2021)

Adam Schwartz, MD, MBA
Mayo Clinic (Associate Professor of Orthopaedic Surgery); Phoenix, AZ
(TEP member 2020 – 2021)

Robert Sterling, MD
Johns Hopkins University School of Medicine (Associate Professor of Orthopaedic Surgery and Vice Chair for Quality, Safety, and Service); Baltimore, MD
(TEP member 2020 – 2021)

Margaret A. VanAmringe, MHS
The Joint Commission (Vice President for Public Policy and Government Relations); Washington, DC
(TEP member 2013 – 2021)

Christine Von Raesfeld
Patient Expert; Santa Clara, CA
(TEP member 2020 – 2021)

Patricia Walker, PhD
Patient Expert; South Holland, IL
(TEP member 2020 – 2021)

Kevin Woodward, PA-C, MMS
American Academy of Physician Assistants, Maryland Academy of Physician Assistants, John Hopkins University (Physician Assistant of Orthopaedic Surgery); Baltimore, MD
(TEP member 2020 – 2021)

Adolph J. Yates, Jr., MD, FAAOS, FAOA
UPMC-Shadyside Hospital (Chief of Orthopedic Surgery), University of Pittsburgh School of Medicine (Professor and Vice Chair for Quality, Department of Orthopedic Surgery); Pittsburgh, PA
(TEP member 2018 – 2021)

Clinical Working Group Members (2020 – 2021):

James I. Huddleston, III, MD
Stanford University Medical Center (Associate Professor and Chief of Arthritis Service, Department of Orthopaedic Surgery), American Association of Hip and Knee Surgeons (Chair of the Advocacy Council); Stanford, CA

Jay R. Lieberman, MD
Keck School of Medicine of the University of Southern California (Professor and Chair, Department of Orthopaedic Surgery, Director of Institute of Orthopaedics), Viterbi School of Engineering of the University of Southern California (Professor of Biomedical Engineering), The Hip Society (Second Vice President); Los Angeles, CA

Mary I. O'Connor, MD
Vori Health (Chief Medical Officer); New Haven, CT

Kathryn Schabel, MD
Oregon Health and Science University (Associate Professor of Orthopaedic Surgery, Adult Reconstruction); Portland, OR

Patient Working Group (2020 –2021):
The Patient Working Group was assembled with the support of Rainmakers, Inc, with the goal of including patient perspectives in decision making during key measure development decisions. The Patient Working Group consists of six patients with experience undergoing THA and TKA procedures. CORE recognizes the value of partnering with patients to discuss the development of this measure.

B. Glossary

Attribution: Assignment of the outcome of a patient episode of care to one or more healthcare provider entities for the purpose of assessing performance quality.

Bootstrapping: The bootstrap is a computer-based method for estimating the standard error of an estimate when the estimate is based on a sample with an unknown probability distribution. Bootstrap methods depend on the bootstrap sample, which is a random sample of size *n* drawn with replacement from the population of *n* objects. The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of the statistic by the empirical standard deviation of the replications.

C-statistic: An indicator of the model's discriminant ability or ability to correctly classify those who have and have not met the SCB improvement following a THA/TKA procedure. Potential values range from 0.5, meaning no better than chance, to 1.0, an indication of perfect prediction. Perfect prediction implies patients' outcomes can be predicted completely by their risk factors, and physicians and hospitals play no role in their patients' outcomes.

Case mix: The particular characteristics of patients with index admissions at a given hospital.

Clinician-specific or clinician group-specific effect: A measure of the clinician or clinician group quality of care that is calculated through hierarchical logistic regression, taking into consideration how many patients were eligible for the cohort, these patients' risk factors, and how many met the SCB improvement outcome. The clinician- or clinician group-specific effect is the calculated random effect intercept for each clinician or clinician group. The clinician- or clinician group-specific intercept will be negative for a better-than-average clinician or clinician group, positive for a worse-than-average clinician or clinician group, and close to zero for an average clinician or clinician group. The clinician- or clinician group-specific intercept is used in the numerator to calculate "predicted" SCB improvement.

Cohort: The defined population included in the measure after inclusion and exclusion criteria have been applied.

Comorbidities: Medical conditions the patient had in addition to his/her primary reason for admission to the hospital.

Condition Categories (CCs): Groupings of ICD-10-CM diagnosis codes in clinically relevant categories, from the HCCs system.[46, 47] CMS uses the grouping, but not the hierarchical logic of the system, to create risk factor variables.

Confidence interval (CI): A CI is a range of values that describes the uncertainty surrounding an estimate. It is indicated by its endpoints; for example, a 95% CI for the odds ratio (OR) associated with protein-calorie malnutrition noted as "1.09 – 1.15" would indicate that there is 95% confidence that the OR lies between 1.09 and 1.15.

Eligible clinician (EC): MIPs evaluates eligibility based on assessment of clinician type, enrollment as a Medicare provider before the performance period, not a qualifying alternative payment model participant (QP) and exceeding the low-volume threshold. An individual MIPS EC is identified through their unique National Provider Identifier (NPI) and Taxpayer Identification Number (TIN) combination and is listed on patient claims. Most NPIs are associated with only one TIN.

Eligible clinician group (EC group): MIPS evaluates eligibility based on assessment of clinician type, enrollment as a Medicare provider before the performance period, not a QP, and being associated with a practice that exceeds the low volume thresholds. An EC group is the aggregate of clinicians within a TIN.

Expected number of procedures with SCB improvement: The number of procedures with an SCB improvement expected based on average clinician's or clinician group's performance with a given clinician's or clinician group's case mix.

Hierarchical model: A widely accepted statistical method that enables evaluation of relative clinician or clinician group performance by accounting for patient risk factors and the number of patients clinicians or clinician groups treat. This statistical model accounts for the hierarchical structure of the data (patients clustered within clinicians or clinician groups are assumed to be correlated) and accommodates modeling of the association between outcomes and patient characteristics. Based on the hierarchical model, we can evaluate (1) how much variation in clinician and clinician group improvement rates overall is accounted for by patients' individual risk factors (such as age and other medical conditions), and (2) how much variation is accounted for by clinician or clinician group contribution to the improvement outcome.

Medicare fee-for-service (FFS): Original Medicare plan in which providers receive a fee or payment for each individual service provided directly from Medicare. Only beneficiaries in Medicare FFS, not in managed care (Medicare Advantage), are included in the measure.

National observed improvement rate: All included hospitalizations with the outcome divided by all included hospitalizations.

Odds ratio (OR): The ORs express the relative odds of the outcome for each of the predictor variables. For example, the OR for Protein-calorie malnutrition (CC 21) represents the odds of the outcome for patients with that risk variable present relative to those without the risk variable present. The model coefficient for each risk variable is the log (odds) for that variable.

Outcome: The result of a broad set of healthcare activities that affect patients' well-being. For the MIPS THA/TKA PRO-PM, the outcome is SCB improvement defined as meeting or exceeding the SCB threshold of 22 points on the HOOS, JR for hip patients and meeting or exceeding the SCB threshold of 20 points on the KOOS, JR for knee patients.

Patient-Reported Outcome (PRO): The concept of a patient-reported outcome.

Patient-Reported Outcome Measure (PROM): The survey instrument that captures patient-reported outcome data. In this measure, the HOOS, JR and KOOS, JR are used for THA and TKA procedures, respectively.

Patient-Reported Outcome-Based Performance Measure (PRO-PM): A performance measure that uses patient-reported outcome data to define the measure outcome (such as this measure).

Predicted number of procedures with SCB improvement: The number of procedures with an improvement outcome predicted based on the clinician's or clinician group's performance with its observed case mix.

Predictive ability: An indicator of the model's discriminant ability or ability to distinguish high-risk subjects from low-risk subjects. A wide range between the lowest decile and highest decile suggests better discrimination.

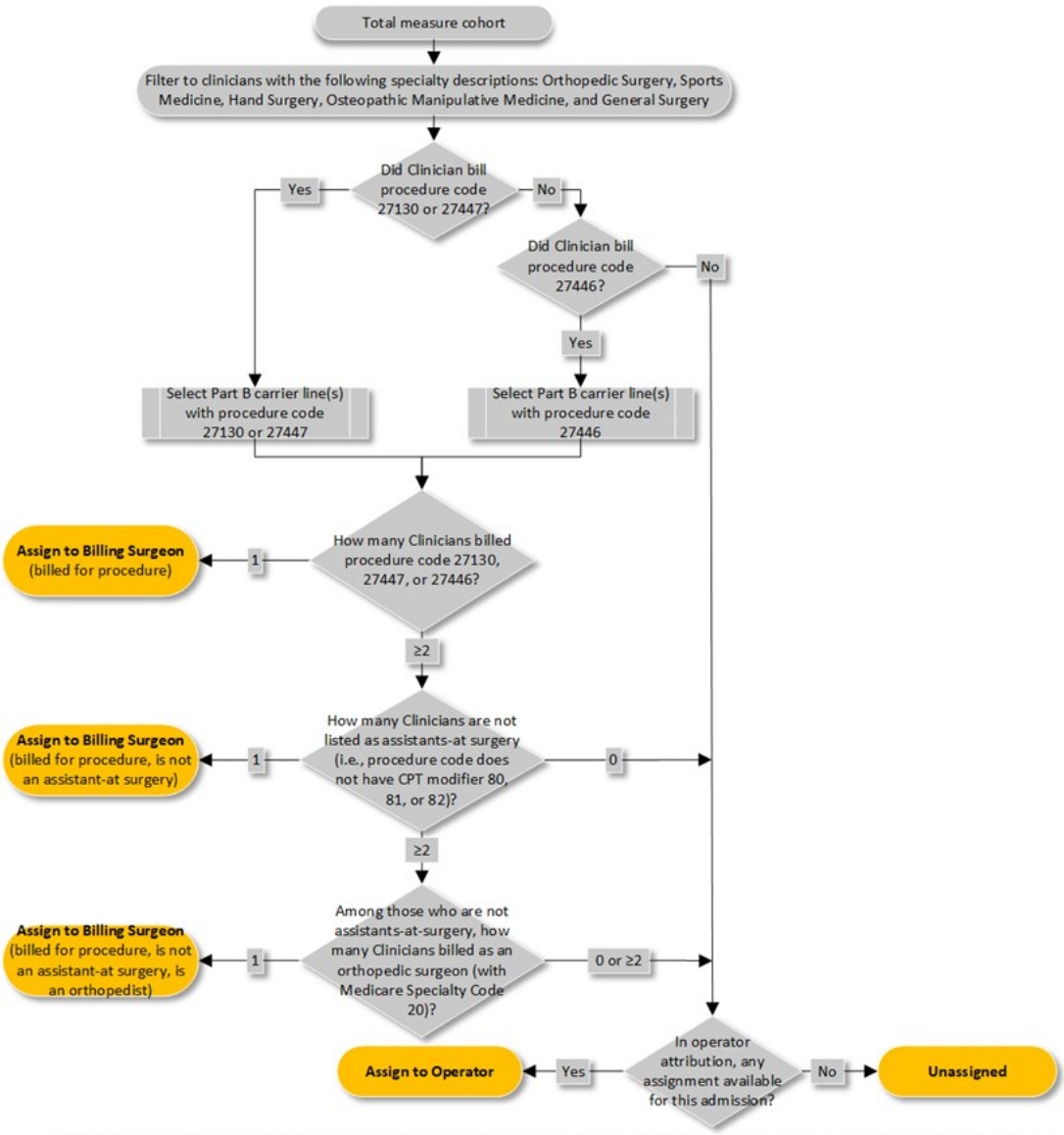
Risk-adjustment variables: Patient demographics and comorbidities used to standardize rates for differences in case mix across clinicians or clinician groups.

C. Attribution

The MIPS THA/TKA PRO-PM uses the attribution approach developed by the MIPS THA/TKA Complication measure. This approach attributes the outcome for each patient in the cohort to a single clinician or clinician group. Figure C1 shows the approach to attribution to the Billing Surgeon at the MIPS EC level. An EC is identified through his/her unique NPI and TIN combination. The MIPS EC group assigned is identified as the TIN on the Part B carrier claim line item for the procedure identified by the attribution in Figure C1 (of note, this figure was developed by the MIPS THA/TKA Complication measure development team).

During measure testing of the MIPS THA/TKA PRO-PM, we identified that the first step of the attribution methodology was assigning a few patients to clinicians not primarily responsible for the THA/TKA procedures, such as Physician Assistants or clinicians with specialty descriptions outside of orthopedics (such as obstetrics/gynecology). Although this did not happen often (N=12, 0.94% of measure cohort), the patient assignment was incongruous with the intention of the attribution methodology. We, therefore, propose the measure attribution only consider the following clinician specialties: orthopedic surgery, sports medicine, hand surgery, osteopathic manipulative medicine, and general surgery. Specifically, we recommend first limiting the Part B claim lines to those with the above clinician specialties and applying the original attribution algorithm afterwards. This refinement will help increase the face validity of the attribution methodology. Our team's orthopedic expert and TEP agreed with the refinement to the attribution methodology.

Figure C1. Approach to Identifying Billing Surgeon in Medicare Claims Data at MIPS EC level; EC group is the TIN of the Attributed EC



KEY

Clinician: A unique National Provider Identifier (NPI) and Taxpayer Identification Number (TIN) combination

Current Procedural Terminology (CPT®) codes

- **27130:** Arthroplasty, acetabular and proximal femoral prosthetic replacement (total hip arthroplasty), with or without autograft or allograft
- **27446:** Arthroplasty, knee, condyle and plateau; medial OR lateral compartment
- **27447:** Arthroplasty, knee, condyle and plateau; medial AND lateral compartments with or without patella resurfacing (total knee arthroplasty)

CPT® modifiers

- **80:** Assistant Surgeon
- **81:** Minimum Assistant Surgeon
- **82:** Assistant Surgeon (when qualified Resident Surgeon not available)

Clinician specialty codes

- **02:** General Surgery
- **12:** Osteopathic Manipulative Therapy
- **20:** Orthopedic Surgery
- **23:** Sport Medicine
- **40:** Hand Surgery

D. Cohort Criteria

Codes used to define elective, primary THA/TKA procedures can be found in the 2020 THA/TKA Complication measure supplemental file available on QualityNet (https://qualitynet.cms.gov/inpatient/measures/complication/resources#tab3).[30]

E. Statistical Approach (Hierarchical Generalized Linear Model [HGLM] Equations)

The MIPS THA/TKA PRO-PM uses an HGLM to estimate RSIRs for ECs and EC groups (providers). This modeling approach accounts for the within-provider correlation of the observed outcome and accommodates the assumption that underlying differences in quality across ECs or EC groups lead to systematic differences in outcomes.

In the MIPS THA/TKA PRO-PM, an HGLM model is estimated. Then for each EC or EC group, a standardized risk ratio (SRR) is calculated. The RSIR is calculated by multiplying the SRR for each EC or EC group by the national observed improvement rate.

F. PROM Surveys and Other Data Elements Collected from Patients

F.1 Preoperative Assessments (to Be Collected between 90 and Zero Days Prior to THA/TKA Procedure)

- Procedure Type
- PROMIS-Global (all items) or VR-12 (all items)
- For THA Patients, HOOS, JR
- For TKA patients, KOOS, JR
- Medicare Provider Number
- Medicare Health Insurance Claim (HIC) Number/ Medicare Beneficiary Identifier (MBI)
- Date of Birth
- Date of Collection
- Mode of Collection
- Person Completing the Survey
- Single-Item Health Literacy Screening (SILS2) Questionnaire
- BMI or Weight (kg)/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)

F.2 Postoperative Assessments (Testing Data Collected between 270 and 365 Days Following THA/TKA Procedure in CJR; We Propose Data Be Collected between 300 and 425 Days Following THA/TKA Procedure for Future Measure Use)

- Procedure Type
- For TKA Patients, KOOS, JR
- For THA Patients, HOOS, JR
- Medicare Provider Number
- HIC Number/ MBI
- Date of Birth
- Date of Collection
- Mode of Collection
- Person Completing the Survey
- Date of Admission to Anchor Hospitalization
- Date of Eligible Procedure

Table F1. Data Elements Used in Measure Testing

Type of Element	Data Element	Format	Range	Collection	Source
Procedure Related	Procedure Type	1 = Left Hip Replacement 2 = Right Hip Replacement 3 = Left Knee Replacement 4 = Right Knee Replacement	1-4	Pre- and postoperative	Provider Reported
	Date of Eligible Procedure	Date (MM/DD/YYYY)	N/A	Postoperative	Medical record/EHR
PROMs	VR-12 (all items)	See Table F2	See Table F2	Preoperative	Patient reported
	PROMIS-Global (all items)	See Table F3	See Table F3	Preoperative	Patient reported
	HOOS, JR (six items)	See Table F4	See Table F4	Pre- and postoperative	Patient reported
	KOOS, JR (seven items)	See Table F5	See Table F5	Pre- and postoperative	Patient reported
PROM related	Date of Collection	Date (MM/DD/YYYY)	N/A	Pre- and postoperative	Provider reported
	Mode of Collection	0 = Paper 1 = Telephone (interactive voice response) 2 = Electronic (web-based, patient portal, Electronic Health Record (EHR), etc.)	0,1,2	Pre- and postoperative	Provider reported

Type of Element	Data Element	Format	Range	Collection	Source
	Person Completing the Survey	0 = Self 1 = Surrogate	0,1	Pre- and postoperative	Patient or provider reported
Elements related to Matching	Date of Birth	Date (MM/DD/YYYY)	N/A	Pre- and postoperative	Provider reported
	Date of Admission to Anchor Hospitalization	Date (MM/DD/YYYY)	N/A	Postoperative	Medical record/EHR
	Medicare Provider Number	Six-digit Medicare provider number, also known as CCN	N/A	Pre- and postoperative	Provider reported
	Medicare Health Insurance Claim (HIC) Number/MBI	Ten or 11-digit account number (e.g., 123456789A)	N/A	Pre- and postoperative	Provider reported
Risk Variables	SILS2 questionnaire ("How comfortable are you filling out medical forms by yourself?")	0 = Not at all 1 = A little bit 2 = Somewhat 3 = Quite a bit 4 = Extremely	0–4	Preoperative	Patient reported
	BMI[a]	Weight (kg)/Height (cm)	10–70	Preoperative	Medical record/EHR
	Height[b]	Centimeters (cm)	Positive number with 1 decimal digit	Preoperative	Medical record/EHR
	Weight[b]	Kilograms (kg)	Positive number with 1 decimal digit	Preoperative	Medical record/EHR
	Use of Chronic (≥ 90 days) Narcotics	0 = No 1 = Yes	0,1	Preoperative	Medical record/EHR (provider reported)

Type of Element	Data Element	Format	Range	Collection	Source
	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?")	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Preoperative	Patient reported
	Quantified Spinal Pain: Patient-Reported Back Pain, Oswestry Index Question ("My BACK PAIN at the moment is")	0 = None 1 = Very mild 2 = Moderate 3 = Fairly severe 4 = Very severe 5 = Worst imaginable	0–5	Preoperative	Patient reported

[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.

Table F2. Data Elements from the VR-12 Health Survey

Item	Format	Range	Collection
General health	1 = Excellent 2 = Very good 3 = Good 4 = Fair 5 = Poor	1–5	Preoperative
Does your health limit you in moderate activities such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	1–3	Preoperative
Does your health limit you in climbing several flights of stairs?	1 = Yes, limited a lot 2 = Yes, limited a little 3 = No, not limited at all	1–3	Preoperative
During the past four weeks, have you accomplished less in work or other daily activities than you would like because of your physical health?	1 = No, none of the time 2 = Yes, a little of the time 3 = Yes, some of the time 4 = Yes, most of the time 5 = Yes, all of the time	1–5	Preoperative
During the past four weeks, were you limited in the kind of work or other daily activities because of your physical health?	1 = No, none of the time 2 = Yes, a little of the time 3 = Yes, some of the time 4 = Yes, most of the time 5 = Yes, all of the time	1–5	Preoperative
During the past four weeks, have you accomplished less in work or other daily activities than you would like as a result of any emotional problems (such as feeling depressed or anxious)?	1 = No, none of the time 2 = Yes, a little of the time 3 = Yes, some of the time 4 = Yes, most of the time 5 = Yes, all of the time	1–5	Preoperative
During the past four weeks, did you not do work or other activities as carefully as usual as a result of any emotional problems (such as feeling depressed or anxious)?	1 = No, none of the time 2 = Yes, a little of the time 3 = Yes, some of the time 4 = Yes, most of the time 5 = Yes, all of the time	1–5	Preoperative
During the past four weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?	1 = Not at all 2 = A little bit 3 = Moderately 4 = Quite a bit 5 = Extremely	1–5	Preoperative

Item	Format	Range	Collection
How much of the time during the past four weeks have you felt calm and peaceful?	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	1–6	Preoperative
How much of the time during the past four weeks have you had a lot of energy?	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	1–6	Preoperative
How much of the time during the past four weeks have you felt downhearted and blue?	1 = All of the time 2 = Most of the time 3 = A good bit of the time 4 = Some of the time 5 = A little of the time 6 = None of the time	1–6	Preoperative
During the past four weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?	1 = All of the time 2 = Most of the time 3 = Some of the time 4 = A little of the time 5 = None of the time	1–5	Preoperative
General physical health now compared to one year ago	1 = Much better 2 = Slightly better 3 = About the same 4 = Slightly worse 5 = Much worse	1–5	Preoperative
Emotional problems (such as feeling anxious, depressed, or irritable) now compared to one year ago	1 = Much better 2 = Slightly better 3 = About the same 4 = Slightly worse 5 = Much worse	1–5	Preoperative

Table F3. Data Elements from PROMIS-Global (Format Reflects PROMIS-Global Version 1.2)

Item	Format	Range	Collection
General health	1 = Poor 2 = Fair 3 = Good	1–5	Preoperative

Item	Format	Range	Collection
	4 = Very good 5 = Excellent		
General quality of life	1 = Poor 2 = Fair 3 = Good 4 = Very good 5 = Excellent	1–5	Preoperative
General physical health	1 = Poor 2 = Fair 3 = Good 4 = Very good 5 = Excellent	1–5	Preoperative
General mental health	1 = Poor 2 = Fair 3 = Good 4 = Very good 5 = Excellent	1–5	Preoperative
General satisfaction with social activities and relationships	1 = Poor 2 = Fair 3 = Good 4 = Very good 5 = Excellent	1–5	Preoperative
General ability to carry out social activities and roles	1 = Poor 2 = Fair 3 = Good 4 = Very good 5 = Excellent	1–5	Preoperative
Ability to carry out every day physical activities	1 = Not at all 2 = A little 3 = Moderately 4 = Mostly 5 = Completely	1–5	Preoperative
Emotional problems in past seven days	1 = Always 2 = Often 3 = Sometimes 4 = Rarely 5 = Never	1–5	Preoperative
Average fatigue in past seven days	1 = Very severe 2 = Severe 3 = Moderate 4 = Mild 5 = None	1–5	Preoperative

Item	Format	Range	Collection
Average pain in past seven days	Ten-point scale (no pain to worst imaginable pain)	0–10	Preoperative

Table F4. Data Elements from HOOS, JR Survey

Data Element	Format	Range	Collection
Pain: Amount of hip pain last week going up or down stairs	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Pain: Amount of hip pain last week walking on an uneven surface	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty last week due to your hip when rising from sitting	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty last week due to your hip when bending to the floor/picking up an object	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty last week due to your hip when lying in bed (turning over, maintaining hip position)	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty last week due to your hip when sitting	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative

Table F5. Data Elements from the KOOS, JR Survey

Data Element	Format	Range	Collection
Stiffness: Severity of knee joint stiffness after first wakening in the morning during the last week	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Pain: Amount of knee pain last week when twisting/pivoting on knee	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Pain: Amount of knee pain last week when straightening knee fully	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Pain: Amount of knee pain last week when going up or down stairs	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Pain: Amount of knee pain last week when standing upright	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty rising from sitting in last week due to knee	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative
Function (Daily Living): Degree of difficulty bending to floor/picking up an object in last week due to knee	0 = None 1 = Mild 2 = Moderate 3 = Severe 4 = Extreme	0–4	Pre- and postoperative