NATIONAL QUALITY FORUM
Driving measurable health improvements together

MEASURE WORKSHEET

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

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

Brief Measure Information

NQF #: 2958

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

Measure Steward: Massachusetts General Hospital

Brief Description of Measure: The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60% or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision. The target population is adult patients who had a primary hip or knee replacement surgery for treatment of hip or knee osteoarthritis.

Developer Rationale: Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

Numerator Statement: The numerator is the number of respondents who have an adequate knowledge score (60% or greater) and a clear preference for surgery.

Denominator Statement: The denominator includes the number of respondents from the target population who have undergone primary knee or hip replacement surgery for treatment of knee or hip osteoarthritis.

Denominator Exclusions: Respondents who are missing 3 or more knowledge items do not get a total knowledge score and are excluded. Similarly, respondents who do not indicate a preferred treatment are excluded. No other exclusions as long as the respondent has the procedure for the designated condition.

Measure Type: Outcome: PRO-PM

Data Source:

Instrument-Based Data

Level of Analysis:

Clinician: Group/Practice

Original Endorsement Date: 10/25/2016

Most Recent Endorsement Date: 10/25/2016 5:

Preliminary Analysis: Maintenance of Endorsement

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

Criteria 1: Importance to Measure and Report

1a. Evidence

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

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

The developer provides the following description for this measure:

- This is a maintenance Patient-Reported Outcome Performance Measure (PRO-PM) at the clinician group/practice level that is derived from patient responses to the Hip or Knee Decision Quality Instruments. It assess the proportion of participants who have a passing knowledge score (60 percent or higher) and a clear preference for surgery. These are considered to have met the criteria for an informed, patient-centered decision. The target population is adult patients who had a primary hip or knee replacement surgery for treatment of hip or knee osteoarthritis.
- The developer states the purpose of engaging patients in decisions is to ensure that they are well informed and received their preferred treatment and that this measure directly assesses the extent to which patients are informed and have a clear preference for hip or knee replacement surgery.

Summary of prior review in 2017:

- During the 2017 measure evaluation meeting, the Person- and Family-Centered Care Standing Committee agreed that asking a patient simple questions such as which treatment do they prefer, do they prefer to have surgery/non-surgical options, etc. should be standard for someone who is actually going to have surgery and if they are not given those options, then they should not be operated on.
- Hip and knee replacements are very common, and the committee agreed that just because a patient is clinically eligible for one of these procedures, does not mean it is the best choice of treatment. Thus, patients who elect to have one of these procedures should be well informed about the risks and benefits and have a clear preference.

Changes to evidence from the last review

☐ The developer attests that there have been no changes in the evidence since the measure was last evaluated.
☒ The developer provided updated evidence for this measure:

- The developer cites:
  - A systematic review that found that informed, patient-centered decisions are associated with higher shared decision making scores
  - A cross-sectional survey conducted at four hospitals affiliated with a large health system that found informed, patient-centered decisions were associated with better physical health and physical function outcomes for patients who had total hip or knee replacement surgery
  - A cluster randomized trial of decision support that found that IPC decisions predicted better outcomes following knee replacement surgery.

Question for the Standing Committee:

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

Guidance From the Evidence Algorithm

Measure is a PRO-PM (box 1)-> Relationship between PRO-PM and at least one healthcare action demonstrated (Box 2)-> Pass

Preliminary rating for evidence: ☒ Pass ☐ No Pass

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

Maintenance measures – increased emphasis on gap and variation

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

- The developer reports data from 3 new data sets (referred to as Sample, 3, 4, and 5).
  - Sample 3 includes 3470 patients who completed the items as part of the Orthopedic Patient Reported Outcomes Measurement system at a large health system from 2018-2022 and come from four sites (two academic medical centers and two community hospitals) with 53 arthroplasty surgeons.
    - The mean IPC rate was 76.5 percent and individual site IPC scores ranged from 72 percent to 80 percent ($p<0.006$)
  - Sample 4 includes data collected from 2016 to 2018 from three sites with 8 surgeons and 559 patients who participated in a randomized trial comparing two different decision aids.
    - All patients received a decision aid about hip or knee replacement surgery as part of their care and completed a survey shortly after the visit with the surgeon and again about 6 months post-operatively.
    - The developer found the overall rate was 92 percent and the range was91 percent-95 percent across sites.
  - Sample 5 includes data from four sites, 22 surgeons and 405 patients who provided sufficient data to calculate an IPC score.
    - Patients were surveyed by mail about 6 months after their surgery.
    - Overall, IPC was 70 percent, and the rates ranged from 62 percent to 77 percent by site.
    - While all sites had access to patient decision aids, a minority of patients (16 percent overall) received a decision aid as part of their care.

Disparities

- The developer examined the three Samples and found significant differences in IPC percent by:
  - Age in Sample 3 (<65 74 percent v. 78 percent, p=0.004)
  - Gender in Sample 3 (Female 75 percent v. Male 79 percent, p=0.003) and Sample 5 (Female 65 percent v. Male 77 percent, p=0.02).
  - Race (White, non-Hispanic 93 percent v. Other Race/ethnicity 82 percent, p=0.04) and education (College degree or more 94.5 percent v. Less than college degree 88 percent, p=0.01) in Sample 4.

Questions for the Standing Committee:

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

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

Criteria 2: Scientific Acceptability of Measure Properties

Complex measure evaluated by the Scientific Methods Panel (SMP)? ☒ **Yes** ☐ **No**

Evaluators: Daniel Deutscher, Dave Nerenz, Eric Weinhandl, Jeff Geppert, Jennifer Perloff, Joe Kunisch, John Bott, Patrick Romano, Paul Kurlansky, Ron Walters, ZQ Lin

- The SMP passed on Reliability with a score of: H-6; M-2; L-0; I-1
- The SMP passed on Validity with a score of: H-4; M-4; L-1; I-0

2a. Reliability: Specifications and Testing

For maintenance measures—no change in emphasis—specifications should be evaluated the same as with new measures.

2a1. Specifications require the measure, as specified, to produce consistent (i.e., reliable) and credible (i.e., valid) results about the quality of care when implemented.

2a2. Reliability testing demonstrates whether the measure data elements are repeatable and producing the same results a high proportion of the time when assessed in the same population during the same time period, and/or whether the measure score is precise enough to distinguish differences in performance across providers.

Specifications:

- Have the measure specifications changed since the last review? ☐ Yes ☒ No
- Measure specifications are clear and precise.
- Measure specifications for the instrument-based measure also include the specific instrument (e.g., PROM[s]); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and the calculation of response rates to be reported with the performance measure results.

Reliability Testing:

- Did the developer conduct new reliability testing? ☒ Yes ☐ No
- Reliability testing was conducted at the accountable-entity level:

- For the current submission, the developer divided data within each practice site into samples with a minimum size of 50. The percentage with IPC within each sample was calculated.
- The reliability was calculated as variability from site divided by total variability. The developer reported that for four groups (site 1 had 16 samples, site 2 had 26 samples, site 3 had 26 samples, and site 4 had four samples), the reliability was 0.735. In the 2016 submission, the developer found that for 14 groups (site 1 had two samples, site 2 had seven samples, site 3 had two samples, and site 4 had three samples), the reliability was 0.853.
- The developer noted that the reliability estimate is slightly lower than the prior submission due to the randomization of individuals to groups.

- Reliability testing was conducted at the patient/encounter level:
  - In the 2016 submission, the developer conducted test-retest reliability of the knowledge and preference items from the same individuals four to six weeks apart.
  - For the knowledge score, the developer examined the ICC of the knowledge score at time #1 and time #2.
  - For the preference item, the developer examined the kappa between the response at time #1 and response at time #2.
  - The test-retest reliability of the knowledge score was examined in sample #1 with an ICC of 0.81 (95 percent CI ranging from 0.71-0.87). The test-retest reliability of the item assessing preferred treatment had a Kappa of 0.801.

SMP Summary:

- The SMP voted to approve the measure on reliability in the pre-meeting vote and did not choose to pull the measure for discussion.

Questions for the Standing Committee regarding reliability:

- *Do you have any concerns that the measure cannot be consistently implemented (i.e., are the measure specifications adequate)?*
- *The SMP is satisfied with the reliability testing for the measure. Does the Standing Committee think there is a need to discuss and/or vote on reliability?*

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

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

For maintenance measures – less emphasis if no new testing data are provided

2b1. Measure Intent: The measure specifications are consistent with the measure's intent and capture the most inclusive target population.

2b2. Validity testing should demonstrate the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

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

Validity Testing

- Did the developer conduct new validity testing? ☒ **Yes** ☐ **No**

- Validity testing was conducted at the accountable-entity level:
    - For the current submission, the developer conducted predictive validity of the overall IPC surgery measure.
    - The developer hypothesized that patients who made IPC decisions would have more engagement in decisions (as measured by the Shared Decision Making [SDM] Process scale); higher confidence (as measured by the SURE [Sure of myself, Understand information, Risk-benefit ratio, and Encouragement]) scale, a short form of the decisional conflict scale); higher satisfaction; and less regret.
    - The developer used generalized linear and logistic regression models with the General Estimating Equations approach to account for clustering of patients within clinicians.
    - The models were adjusted for patient age, gender, education, joint, and baseline quality of life scores.
        - For hip and knee surgery decisions, the developer found IPC was significantly associated with higher shared decision making scores (mean SDM Process = 2.3 for non-IPC versus 2.7 IPC group, p<0.001) and higher decision confidence (SURE top score = 63 percent for non-IPC versus 92.3 percent IPC group, p<0.001).
        - Controlling for age, sex, surgical status, education, and diagnosis (osteoarthritis versus spine), the developer found participants who made IPC decisions were more likely to be extremely satisfied with their pain (odds ratio [OR] of 2.45; 95 percent CI of 1.45–4.15; and P = 0.0008), were more likely to be very or extremely satisfied with their treatment (an OR of 2.59; 95 percent CI of 1.59–4.22; and P = 0.0001), and reported less regret (–5.63 points; 95 percent CI of –8.25 to –3.01; and P = 0.0001) than those who did not make IPC decisions.
    - The developer also tested hypotheses that IPC surgery is associated with better health outcomes using a linear regression model with quality of life at six months post-surgery as the dependent variable and IPC, age, education, sex, treatment (surgery versus nonsurgery), joint (hip versus knee), site, and baseline quality of life (SF-12 physical component score) as independent variables.
        - The developer found that the IPC was significantly associated with improvements in overall (0.05 points [Standard Error of the Mean (SE) 0.02] for EuroQol-5 Dimension (EQ-5D), p=0.004) and disease-specific quality of life (4.22 points [SE 1.82] for knee p=0.02, and 4.46 points [SE 1.54] for hip, p=0.004).
        - The developer stated that the IPC was related to overall (mean difference EQ-5D 0.04 points [0.02, 0.07], p<0.001) and disease-specific quality of life (mean difference 4.9 points [1.5, 8.3], p=0.004) for knee but not hip patients.

- Validity testing was conducted at the patient/encounter level:
    - For the 2016 submission, the developer developer performed discriminant validity of the knowledge assessment by comparing scores of those who should have higher knowledge (e.g., scores of patients who had used a decision aid versus those who did not).
    - The developer stated that the mean knowledge scores discriminated between patients in a decision aid group with 67 percent (SD of 21.2) compared to 51 percent (SD of 24.9) in the usual care group (p<0.001).

Exclusions

- The developer states that respondents who skip three or more knowledge items or the preference item do not receive a total score.

- The developer states that for the current submission, it did not find significant or meaningful differences by site or patient characteristics due to exclusions.
- In sample 5, gender was significant in one sample (suggesting females were more likely to have missing data), but the numbers were small, and the developer did not find a similar result in sample 4 (in which females were less likely to have missing data).

Risk Adjustment

- The measure is not risk-adjusted or stratified.
- The developer states that it does not recommend risk adjustment for this measure. Any patient who has one of these elective surgeries should be able to answer the knowledge questions correctly and should have a clear preference for the procedure (to meet the standards of informed consent).

Meaningful Differences

- For the current submission, the developer notes data from one health system (sample 3) that has been focused on shared decision making and has decision aids available for patients, which suggests that sites can achieve rates in the 70–80 percent range.
- The developers also cite the DECIDE Osteoarthritis (DECIDE-OA) trial (sample 4), which achieved rates of IPC at the three sites (> 90 percent).

Missing Data

- The developer reports that missingness is small for both sample 4 (9/568 [1.6 percent]) and for sample 5 (13/405 [3 percent]). The developer notes that patient characteristics (e.g., age, gender, and race/ethnicity) did not vary significantly between those who had and did not have missing data.

Comparability

The measure only uses one set of specifications for this measure

SMP Summary:

- The SMP voted to pass the measure on validity in the pre-meeting and did not pull the measure for discussion. In the written comments, there were a few concerns raised about validity.
- One member found the accountable entity level testing to be limited to date, but the results still support validity.
- Regarding missing data, one reviewer noted that there were too few respondents but most reviewers wrote that the level of missing data was in line with other survey-based measures.
- One SMP member noted that there may be some concerns in the future when and if the measure is used in patient populations with lower education or literacy levels.
- SMP members had some concerns about risk adjustment, as the developer did not recommend adjusting the measure despite finding a significant effect of the SF-12 score. Some members also noted that the developer did not provide sufficient conceptual rationale for the lack of risk adjustment.
- One SMP member suggested that the developers add additional explanation of why 60 percent was chosen as the threshold for knowledgeable or unknowledgeable.

Questions for the Standing Committee regarding validity:

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

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

Criterion 3. Feasibility

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

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

- The patient-report surveys can be administered online to support electronic capture via patient reported outcomes registries or other online survey platforms. If administered via mail or paper, then it will require staff at sites to enter the patient data into an online database for analysis.
- The developer reports that at one health system, the items have been incorporated into the Patient-Reported Outcomes registry and are captured and scored as part of routine orthopedic care for patients undergoing surgery for hip, knee and spine conditions.
- The developer reports that the administration of these questions has been conducted across multiple sites, in multiple modes (predominantly paper and online surveys). A large health system has incorporated the items into their patient-reported outcomes registry for orthopedics and the data is being collected as part of routine care in that system. Generally, the developer states, patients find these surveys acceptable as indicated by good response rates and low missing data. However, whether administered as a stand-alone survey or as part of a patient-reported outcomes measure set, to obtain sufficiently high response rates often requires effort on the part of clinic staff (for example to remind patients to complete). Further, as mentioned in prior submission response below, it is easier to identify and survey patients who undergo surgery than those who pursue non-operative care.
- There are no fees for the measure or for the use of the Hip or Knee Decision Quality Instruments used to generate the measure, provided the surveys are used in accordance with the creative commons copyright license.

Questions for the Standing Committee:

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

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

Criterion 4: Use and Usability

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

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

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

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

Current uses of the measure

Publicly reported?	☐	Yes	☒	No		
Current use in an accountability program?	☒	Yes	☐	No	☐	UNCLEAR
Planned use in an accountability program?	☐	Yes	☐	No	☒	NA

Accountability program details

- The developer states the measure is used in the following programs:
    - Blue Cross Blue Shield of Massachusetts Alternative Quality Contract – They are currently piloting it and by the end of 2022 are planning to collect data directly from their members using these instruments as a basis for confidential reporting to providers, with the goal of using these performance data as a basis for financial incentives.
    - The Alliance Quality Path Program specifies measurement of decision quality and shared decision making as part of their criteria for recognition. NQF #2958 can be used for this recognition.
    - Shared Decision Making Program at Massachusetts General Brigham Health System incorporates items IPC measure into the Patient Reported Outcomes Registry. Responses are summarized across surgeons and practices, used to identify high and low performing clinicians, and used to promote quality improvement initiatives in the departments. The initiative is also working to integrate patient decision aids into routine orthopedic care. Massachusetts Aligned Measure Set for Global Budget-Based Risk Contracts sponsored by Executive Office of Health and Human Services (EOHHS), the Massachusetts Health Policy Commission (HPC), and the Center for Health Information Analysis (CHIA). The IPC measure 2958 is part of the aligned measure set that is available for use in payment programs in Massachusetts. It is not part of the core set, and as a result, it is not mandatory. Rather it is on the 'menu set' and available for use. At this time, we do not know whether any health systems, hospitals or other entities have selected to use this as part of their measures or whether any other insurers (aside from BCBS MA as described above) have incorporated the measures into their contracts.

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

Feedback on the measure provided by those being measured or others

- The developer reports having feedback from patients who have participated in research studies using these measures. They report that the measures are highly acceptable to patients with very little missing data. The developer has heard that patients are interested in the correct answers to the knowledge items and, when possible, we make those available after the assessment is completed. When we have shared results with the surgeons, we have had generally positive feedback. They often want to see the item-level responses to understand knowledge gaps or areas where patients have

misperceptions that may be driving the scores and/or differences in the scores. The individual knowledge item results will identify areas where patients consistently have inaccurate understanding about options, benefits and harms. Occasionally, surgeons have challenged whether a particular knowledge answer is "correct." The developer shares the annotated evidence-base used to support the correct and incorrect responses. If they have new evidence, then the developer will consider changing the items and/or responses to reflect updated evidence. This open and transparent process often leads to them accepting the items and results.

- The developer states that they have used the feedback to update the user guide where we provide advice to users on how to best set up the survey to ensure high response rates and high quality data. The main advice has been to incorporate the survey items into existing registries or patient survey platforms supported by electronic medical records. In addition, the developer has advised groups to be prepared to share the correct answers to the knowledge items after the surveys have been completed.

***Questions for the Standing Committee**:*

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

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

4b. Usability (4b1. Improvement; 4b2. Benefits of measure)

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

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

Improvement results

- The developer does not provide trend data but does note that studies using the IPC measure have found patients provided with decision support interventions have significantly higher rates compared to usual care. More recent studies have also shown that when patients receive decision aid as part of routine care, that scores can be quite high (91 percent-95 percent); where as in practices with few patients receiving decision aids, scores are much lower (72-80 percent).

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

Unexpected findings (positive or negative) during implementation

- The developer did not report any unexpected findings during implementation of this measure.

Potential harms

- The developer did not report any potential harms from implementation of this measure.

***Questions for the Standing Committee**:*

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

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

Criterion 5: Related and Competing Measures

Related/Competing Measures

- None

Criteria 1: Importance to Measure and Report

1a. Evidence

1ma.01. Indicate whether there is new evidence about the measure since the most recent maintenance evaluation. If yes, please briefly summarize the new evidence, and ensure you have updated entries in the Evidence section as needed.

[Response Begins]

Yes

[Yes Please Explain]

We have new evidence from three different samples, a randomized controlled trial of decision support, a cross-sectional survey of recent surgical patients, and a prospective sample collected as part of routine care. These data are described in detail in the relevant sections.

[Response Ends]

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

Current Submission:

Updated evidence information here.

Previous (Year) Submission:

Evidence from the previous submission here.

1a.01. Provide a logic model.

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

[Response Begins]

Current Submission:

The purpose of engaging patients in decisions is to ensure that they are well informed and received their preferred treatment (Barry et al 2018). The measure directly assesses the extent to which patients are informed and have a clear preference for hip or knee replacement surgery.

Barry, MJ, Edgman-Levitan, S, Sepucha, K. Shared Decision-Making: Staying focused on the ultimate goal. NEJM Catalyst, 2018 Sep 6. https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0097

Previous Submission:

A high quality decision about elective surgery, such as total hip or knee replacement, requires that patients are well-informed and have a clear preference for surgery. The Informed, Patient Centered (IPC) surgery measure presents data on how well centers or hospitals are doing informing patients and tailoring treatments to patients' preferences.

[Response Ends]

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

Describe how and from whom input was obtained.

[Response Begins]

Current Submission:

Patients and clinicians were involved in the development of the surveys through item generation (n=88 patients and n=51 clinicians), cognitive testing (n=10), and field testing (n=489 patients and n=77 clinicians). Feedback from patients and from clinicians whose patients completed the surveys demonstrates the value of the items. Specifically, patients often ask for answers to the knowledge items, as they are interested in making sure they understand the information and got the items 'correct.' Some even remark that their surgeons did not share this information with them, and they want to know the answers. Further, clinicians have asked to see their patients' responses to determine how much they need to talk about in order to make sure patients are adequately informed. More details can be found in the following articles: Sepucha et al 2008 and Sepucha et al 2011.

Sepucha KR, Levin CA, Uzogara EE, Barry MJ, O'Connor AM, Mulley AG. Developing instruments to measure the quality of decisions: early results for a set of symptom-driven decisions. Patient Educ Couns. 2008 Dec;73(3):504-10. doi: 10.1016/j.pec.2008.07.009. Epub 2008 Aug 20. PMID: 18718734.

Sepucha KR, Stacey D, Clay CF, Chang Y, Cosenza C, Dervin G, Dorrwachter J, Feibelmann S, Katz JN, Kearing SA, Malchau H, Taljaard M, Tomek I, Tugwell P, Levin CA. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. BMC Musculoskelet Disord. 2011 Jul 5;12:149. doi: 10.1186/1471-2474-12-149. PMID: 21729315; PMCID: PMC3146909.

[Response Ends]

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

[Response Begins]

Current Submission:

A recent systematic review demonstrated that informed, patient-centered decisions are associated with higher shared decision making scores (Valentine et al 2021a). A recent cross-sectional survey conducted at four hospitals affiliated with a large health system found the Informed, Patient Centered decisions were associated with better physical health and physical function outcomes for patients who had total hip or knee replacement surgery (Valentine et al 2021b). In addition, a large cluster randomized trial of decision support found that IPC decisions predicted better outcomes following knee replacement surgery (Sepucha et al 2022).

Valentine, KD, Vo H, Fowler FJ Jr, Brodney S, Barry MJ, Sepucha KR. Development and Evaluation of the Shared Decision Making Process Scale: A Short Patient-Reported Measure. Med Decis Making. 2021a Feb;41(2):108-119. doi: 10.1177/0272989X20977878. Epub 2020 Dec 15. PMID: 33319648.

Valentine KD, Cha T, Giardina JC, Marques F, Atlas SJ, Bedair H, Chen AF, Doorly T, Kang J, Leavitt L, Licurse A, O'Brien T, Sequist T, Sepucha K. Assessing the quality of shared decision making for elective orthopedic surgery across a large healthcare system: cross-sectional survey study. BMC Musculoskelet Disord. 2021b Nov 19;22(1):967. doi: 10.1186/s12891-021-04853-x. PMID: 34798866; PMCID: PMC8605511.

Sepucha KR, Vo H, Chang Y, Dorrwachter JM, Dwyer M, Freiberg AA, Talmo CT, Bedair H. Shared Decision-Making Is Associated with Better Outcomes in Patients with Knee But Not Hip Osteoarthritis: The DECIDE-OA Randomized Study. J Bone Joint Surg Am. 2022 Jan 5;104(1):62-69. doi: 10.2106/JBJS.21.00064. PMID: 34437308.

Previous (2016) Submission:

The measure is a PRO that reflects the quality of the treatment decision making process. The measure reflects multiple care processes and outcomes such as communication, provision of information, shared decision making, and patient engagement.

The use of patient decision aids has been associated with increased decision quality. Further, increased decision quality, and having treatments that match patients' preferences, has been associated with reduced utilization of joint replacement surgery and better health outcomes. [Sepucha et al 2011; Sepucha et al 2013; Stacey et al 2014]

1. Sepucha K, Stacey D, Clay C, Chang Y, Cosenza C, Dervin G, Dorrwachter J, Feibelmann S, Katz JN, Kearing S, Malchau H, Taljaard M, Tomek I, Tugwell P, Levin C. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. BMC Musculoskelet Disord 2011 Jul 5;12(1):149.
2. Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.
3. Stacey D, Légaré F, Col N, Bennett C, Barry M, Eden K, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2014 Jan 28(1).

[Response Ends]

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

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

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

[Response Begins]

Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

[Response Ends]

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

Include mean, std dev, min, max, interquartile range, and scores by decile. Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities include. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

Current Submission:

We have new data from 3 new data sets (referred to as Sample, 3, 4, and 5). Sample 3 includes 3470 patients who completed the items as part of the Orthopedic Patient Reported Outcomes Measurement system at a large health system from 2018-2022 and come from four sites (two academic medical centers and two community hospitals) with 53 arthroplasty surgeons. The mean IPC rate was 76.5% and individual site IPC scores ranged from 72% to 80% ($p<0.006$) (see Table 1b.02.a). All sites had access to patient decision aids and have made concerted efforts at shared decision making; however, actual decision aid delivery to surgical patients was quite low across the sites.

Table 1b.02.a: Patient characteristics by site for Sample 3

*	Site 1	Site 2	Site 3	Site 4	Overall
Patient N	1130	990	1134	216	3470
Surgeon N	16	18	18	7	53+
JointHip N(%) (versus knee)	520 (46 %)	496 (50 %)	518 (46 %)	97 (45 %)	1631 (47%)
Patient Age M(SD)	68 (10)	68 (10)	67 (9)	69 (9)	68 (10)
Patient Sex: Female N (%)	651 (58 %)	612 (62 %)	631 (56 %)	125 (58 %)	2019 (58%)
% receiving a decision aid	375 (33%)	117 (12%)	180 (16%)	1 (0.5%)	673 (19%)
IPC score N (%)	843 (74.6%)	749 (75.7%)	907 (80.0%)	156 (72.2%)	2655 (76.5%)

*Cell intentionally left blank;

+Some surgeons operate at more than one site.

Sample 4 includes data collected from 2016 to 2018 from three sites with 8 surgeons and 559 patients who participated in a randomized trial comparing two different decision aids. All patients received a decision aid about hip or knee replacement surgery as part of their care and completed a survey shortly after the visit with the surgeon and again about 6 months post-operatively. Notably, we find very high rates of IPC in the post visit sample, overall rate was 92% and range (91%-95%) across sites, demonstrating that it is possible for sites to obtain high scores.

Table 1b.02.b: Patient characteristics by site for Sample 4:

*	Hospital 1 N=108	Hospital 2 N=165	Hospital 3 N=286	Overall N=559
Surgeon N	2	3	3	8
Hip N (%) versus Knee	49 (45%)	72 (44%)	95 (33%)	216 (39%)
Patient Age M (SD)	66 (10)	64 (9)	65 (9)	65 (9)
Patient Sex: Female N (%)	66 (61)	85 (52)	165 (58)	316 (57)
% patients receiving a decision aid	100%	100%	100%	100%
IPC score N (%)	99 (92%)	156 (95%)	259 (91%)	514 (92%)

*Cell intentionally left blank

Sample 5 includes data from four sites, 22 surgeons and 405 patients who provided sufficient data to calculate an IPC score. Patients were surveyed by mail about 6 months after their surgery. Overall, IPC was 70%, and the rates ranged from (62% to 77%) by site. While all sites had access to patient decision aids, a minority of patients (16% overall) received a decision aid as part of their care.

Table 1b.02.c: Patient characteristics by site for Sample 5.

*	Hospital 1 n=136	Hospital 2 n=130	Hospital 3 n=29	Hospital 4 n=97	Overall n=392
Surgeon n	10	6	4	5	22+
Hip % versus Knee	61%	48%	45%	44%	51%
Patient Age M (SD)	67 (9)	65 (9)	65 (6)	66 (10)	66 (9)
Patient Sex: Female %	59%	56%	48%	46%	54%
% receiving a decision aid	1%	22%	0%	32%	16%
IPC score %	68%	70%	62%	77%	70%

*Cell intentionally left blank

+Surgeons may operate at more than one hospital.

Previous Submission:

The sample includes patients from three sites and a general population sample from the Boston area. The site that had a formal shared decision making process (SDM site) had a higher rate of informed, patient centered (IPC) surgery than the sites with no formal shared decision making (usual care sites). The association between SDM site and rates of IPC surgery remained significant in multivariate analyses controlling for joint (knee/hip), gender, surgery, and decision making process scores [Sepucha et al 2013].

Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002 @. Epub 2013 Jul 25.

[Response Ends]

1b.03. If no or limited performance data on the measure as specified is reported above, then provide a summary of data from the literature that indicates opportunity for improvement or overall less than optimal performance on the specific focus of measurement. Include citations.

[Response Begins]

Current submission:

Published studies have found low rates of Informed, Patient Centered decisions (IPC) in usual care, as well as evidence that interventions, such as decision aids, can increase informed, patient-centered decisions. Jayakumar et al 2021 found that patients randomly assigned to use a decision aid had significantly higher decision quality scores (measured with IPC) than usual care, mean difference 20% SE, 3.02; 95% CI, 14.2%-26.1%; $P < .001$. In a pilot randomized trial, Stacey et al 2014 also found a significant increase in the rates of informed, patient-centered decisions for patients considering knee replacement surgery who received a decision aid compared to usual care (56.4% decision aid arm versus 25.0% usual care; $p < 0.001$). In a larger randomized trial, Stacey et al 2016 found about a 12% difference for those who received a decision aid about knee replacement surgery compared to usual care (56.1% intervention and 44.5% control (Relative risk (RR) 1.25; 95% CI 1.00-1.56, P = 0.05).

Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, Aksan N, Rathouz PJ, Bozic KJ. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults With Knee Osteoarthritis: A Randomized Clinical Trial. JAMA Netw Open. 2021 Feb 1;4(2):e2037107. doi: 10.1001/jamanetworkopen.2020.37107. PMID: 33599773.

Stacey D, Hawker G, Dervin G, Tugwell P, Boland L, Pomey MP, O'Connor AM, Taljaard M. Decision aid for patients considering total knee arthroplasty with preference report for surgeons: a pilot randomized controlled trial. BMC Musculoskelet Disord. 2014 Feb 24;15:54. doi: 10.1186/1471-2474-15-54. PMID: 24564877; PMCID: PMC3937455.

Stacey D, Taljaard M, Dervin G, Tugwell P, O'Connor AM, Pomey MP, Boland L, Beach S, Meltzer D, Hawker G. Impact of patient decision aids on appropriate and timely access to hip or knee arthroplasty for osteoarthritis: a randomized controlled trial. Osteoarthritis Cartilage. 2016 Jan;24(1):99-107. doi: 10.1016/j.joca.2015.07.024. PMID: 26254238.

Previous (2016) Submission:

The DECISIONS study was a national random sample of patients surveyed by telephone up to two years after their decision. They asked earlier versions of four of these knowledge items and found that on the whole, patients had considerable knowledge gaps. For the 141 patients who had discussed hip or knee replacement surgery with their health care provider, the total knowledge score was 32.1% [Fagerlin 2010]. When the researchers combined respondents across different types of elective surgery including back surgery and cataract surgery, race and education were predictors of knowledge (lower education and non White race were associated with lower knowledge).

In summary, data show that patients are not typically well informed about the treatment options for knee and hip replacement surgery, and patients undergo these elective procedures without a clear preference for it. There is considerable room for improvement in elective hip and knee replacement decisions. There is also evidence that clinical sites that have processes in place to promote share decision making (such as use of patient decision aids) are able to achieve higher rates of IPC surgery than the average or usual care.

Fagerlin A, Sepucha K, Couper M, Levin C, Ubel P, Singer E, Zikmund-Fisher B. Patients' knowledge about 9 common health conditions: Data from a national representative sample. Medical Decision Making Sept/Oct 2010 30: 35S-52S, doi:10.1177/0272989X10378700.

[Response Ends]

1b.04. Provide disparities data from the measure as specified (current and over time) by population group, e.g., by race/ethnicity, gender, age, insurance status, socioeconomic status, and/or disability.

Describe the data source including number of measured entities; number of patients; dates of data; if a sample, characteristics of the entities included. Include mean, std dev, min, max, interquartile range, and scores by decile. For measures that show high levels of performance, i.e., "topped out", disparities data may demonstrate an opportunity for improvement/gap in care for certain sub-populations. This information also will be used to address the sub-criterion on improvement (4b) under Usability and Use.

[Response Begins]

Current submission:

The following tables provide data on rates of IPC based on different patient characteristics (including age, gender, race/ethnicity, and education) for the 3 samples that have been described in detail earlier (see 1b.02).

Table 1b.04 Samples 3-5 data on disparities by population

*	*	Sample 3	*	*	Sample 4	*	*	Sample 5	*
Group	N	IPC %	p	N	IPC %	p	N	IPC %	p
Overall	3470	76.5%	*	559	92%	*	392	70%	*
<65	1186	74%	0.004	266	94%	0.22	169	67%	0.22
65+	2284	78%	*	293	90%	*	223	73%	*

*	*	Sample 3	*	*	Sample 4	*	*	Sample 5	*
Female	2019	75%	0.003	316	90%	0.06	212	65%	0.02
Male	1451	79%	*	243	95%	*	180	77%	*
White, non Hispanic	n/a	*	*	515	93%	0.04	369	71%	0.42
Other Race & Ethnicity	n/a	*	*	33	82%	*	23	61%	*
College degree or more	n/a	*	*	347	94.5%	0.01	n/a	*	*
Less than college degree	n/a	*	*	207	88%	*	n/a	*	*

*Cell intentionally left blank

Previous (2016) Submission:

The data come from a sample of patients who were surveyed about one year after surgery or after a visit with an orthopedic surgeon. The covariates we looked at were age (>65, <=65), education (college or more, less than college degree), race/ethnicity (non Hispanic White, other) and gender.

Table: Disparities Data for Knee and Hip Replacement Surgery

VARIABLE GROUP IPC P-value N

EDUCATION >=COLLEGE 57.7% .09 208

<COLLEGE 48.8% 160

RACE NON-HISPANIC WHITE 54.5% .08 352

OTHER RACES 31.2% 16

AGE <65 52.9% .83 153

65+ 54.4% 215

SEX MALE 51.2% .35 165

FEMALE 56.4% 209

JOINT HIP 58.5% .08 176

KNEE 49.0% 198

IPC=informed, patient centered

For the comparison on race/ethnicity, the small number of cases limits the power to detect significant differences.

[Response Ends]

1b.05. If no or limited data on disparities from the measure as specified is reported above, then provide a summary of data from the literature that addresses disparities in care on the specific focus of measurement. Include citations. Not necessary if performance data provided in above.

[Response Begins]

Current Submission:

We found differences by gender in the three samples, as males had higher scores than females. One sample also found that younger respondents had higher scores than older respondents, though the magnitude was relatively small and the other two samples did not find a similar result. The one sample that collected education found that those with college degree had higher rates; however, the magnitude of the difference was modest (~6%) and not likely to be clinically meaningful. Overall, even though statistically significant, the magnitude of these difference were generally small,

suggesting there are not large disparities by these patient characteristics. For the comparison on race/ethnicity, the small number of cases limits the power to detect disparities.

Previous (2016) Submission

Although we did not find significant relationship in this sample between rates of informed, patient-centered surgery and education, there is evidence that less education and non White race are associated with lower knowledge scores (Fagerlin et al, 2010).

Fagerlin A, Sepucha K, Couper M, Levin C, Ubel P, Singer E, Zikmund-Fisher B. Patients' knowledge about 9 common health conditions: Data from a national representative sample. Medical Decision Making Sept/Oct 2010 30: 35S-52S, doi:10.1177/0272989X10378700.

[Response Ends]

Criteria 2: Scientific Acceptability of Measure Properties

2a. Reliability

spma.01. Indicate whether there are changes to the specifications since the last updates/submission. If yes, update the specifications in the Measure Specifications section of the Measure Submission Form, and explain your reasoning for the changes below.

[Response Begins]

No

[Response Ends]

spma.02. Briefly describe any important changes to the measure specifications since the last measure update and provide a rationale.

For annual updates, please explain how the change in specifications affects the measure results. If a material change in specification is identified, data from re-testing of the measure with the new specifications is required for early maintenance review.

For example, specifications may have been updated based on suggestions from a previous NQF CDP review.

[Response Begins]

Not applicable.

[Response Ends]

sp.01. Provide the measure title.

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

[Response Begins]

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

[Response Ends]

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

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

[Response Begins]

The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60% or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision.
The target population is adult patients who had a primary hip or knee replacement surgery for treatment of hip or knee osteoarthritis.

[Response Ends]

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

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

Please do not select:

- *Surgery: General*

[Response Begins]

Musculoskeletal

Musculoskeletal: Joint Surgery

Musculoskeletal: Osteoarthritis

Surgery: Orthopedic

[Response Ends]

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

[Response Begins]

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

Safety

Safety: Overuse

[Response Ends]

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

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

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

Please do not select:

- *Populations at Risk: Populations at Risk*

[Response Begins]

Elderly (Age >= 65)

Populations at Risk: Dual eligible beneficiaries of Medicare and Medicaid

Populations at Risk: Individuals with multiple chronic conditions

Populations at Risk: Veterans

Women

[Response Ends]

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

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

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

Please do not select:

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

[Response Begins]

Clinician: Group/Practice

[Response Ends]

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

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

[Response Begins]

Ambulatory Care

Outpatient Services

[Response Ends]

sp.09. Provide a URL link to a web page specific for this measure that contains current detailed specifications including code lists, risk model details, and supplemental materials.

Do not enter a URL linking to a home page or to general information. If no URL is available, indicate "none available".

[Response Begins]

The survey is available in both English and Spanish and can be accessed at the following website:

https://mghdecisionsciences.org/tools-training/decision-quality-instruments/

[Response Ends]

sp.12. Attach the data dictionary, code table, or value sets (and risk model codes and coefficients when applicable). Excel formats (.xlsx or .csv) are preferred.

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

[Response Begins]

No data dictionary/code table – all information provided in the submission form

[Response Ends]

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

sp.13. State the numerator.

Brief, narrative description of the measure focus or what is being measured about the target population, i.e., cases from the target population with the target process, condition, event, or outcome).

DO NOT include the rationale for the measure.

[Response Begins]

The numerator is the number of respondents who have an adequate knowledge score (60% or greater) and a clear preference for surgery.

[Response Ends]

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

sp.14. Provide details needed to calculate the numerator.

All information required to identify and calculate the cases from the target population with the target process, condition, event, or outcome such as definitions, time period for data collection, specific data collection items/responses, code/value sets.

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

The numerator is the number of respondents who have a positive decision quality assessment.

The numerator is calculated based on patient responses to 6 questions from the Hip or Knee Decision Quality Instruments (these items are listed below in S.18 and included as an appendix): five multiple choice knowledge items and one preference item. One point is awarded for each correct knowledge item and then a total knowledge score is calculated and scaled from (0-100%). Respondents who score 60% or higher on knowledge and who indicate a clear preference for surgery have a positive decision quality assessment and are counted in the numerator. Those who score less than 60% and/or who are either unclear or prefer nonsurgical options have a negative decision quality assessment, and are not counted in the numerator.

[Response Ends]

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

sp.15. State the denominator.

Brief, narrative description of the target population being measured.

[Response Begins]

The denominator includes the number of respondents from the target population who have undergone primary knee or hip replacement surgery for treatment of knee or hip osteoarthritis.

[Response Ends]

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

sp.16. Provide details needed to calculate the denominator.

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

Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

The denominator is all adult patients who had a primary hip or knee replacement surgery for treatment of osteoarthritis and responded to the Hip or Knee Decision Quality Instrument. There is an attached excel file with ICD 10 and CPT codes needed to identify eligible patients to be surveyed for inclusion in the measure. A published manuscript describes the development and validation of an algorithm using ICD 10 and CPT codes that can be used to identify eligible patients to be surveyed for inclusion in the measure (Giardina et al. 2020).

Giardina JC, Cha T, Atlas SJ, Barry MJ, Freiberg AA, Leavitt L, Marques F, Sepucha K. Validation of an electronic coding algorithm to identify the primary indication of orthopedic surgeries from administrative data. BMC Med Inform Decis Mak. 2020 Aug 12;20(1):187. doi: 10.1186/s12911-020-01175-1. PMID: 32787849; PMCID: PMC7425151.

[Response Ends]

sp.17. Describe the denominator exclusions.

Brief narrative description of exclusions from the target population.

[Response Begins]

Respondents who are missing 3 or more knowledge items do not get a total knowledge score and are excluded. Similarly, respondents who do not indicate a preferred treatment are excluded. No other exclusions as long as the respondent has the procedure for the designated condition.

[Response Ends]

sp.18. Provide details needed to calculate the denominator exclusions.

All information required to identify and calculate exclusions from the denominator such as definitions, time period for data collection, specific data collection items/responses, code/value sets – Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format at sp.11.

[Response Begins]

Respondents missing 3, 4, or 5 knowledge responses. Respondents missing a response to the preference item.

[Response Ends]

sp.19. Provide all information required to stratify the measure results, if necessary.

Include the stratification variables, definitions, specific data collection items/responses, code/value sets, and the risk-model covariates and coefficients for the clinically-adjusted version of the measure when appropriate. Note: lists of individual codes with descriptors that exceed 1 page should be provided in an Excel or csv file in required format in the Data Dictionary field.

[Response Begins]

Not applicable.

[Response Ends]

sp.20. Is this measure adjusted for socioeconomic status (SES)?

[Response Begins]

No

[Response Ends]

sp.21. Select the risk adjustment type.

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

[Response Begins]

No risk adjustment or risk stratification

[Response Ends]

sp.22. Select the most relevant type of score.

Attachment: If available, please provide a sample report.

[Response Begins]

Rate/proportion

[Response Ends]

sp.23. Select the appropriate interpretation of the measure score.

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

[Response Begins]

Passing score defines better quality

[Response Ends]

sp.24. Diagram or describe the calculation of the measure score as an ordered sequence of steps.

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

[Response Begins]

The following steps need to be taken to calculate the measure: (1) identify eligible patients (2) administer the Hip or Knee Decision Quality Instrument (3) collect and code responses (4) calculate total knowledge scores and exclude those with 3 or more knowledge items missing (5) calculate the numerator (informed and clear preference for surgery or not) for each individual, excluding those with no knowledge score and/or no preference item and (6) aggregate the measure into a rate over the center or practice.

Responses to five knowledge questions and one preference item from the Hip or Knee Decision Quality Instrument are needed to calculate the Informed, Patient Centered (IPC) surgery measure and are coded and scored as indicated below.

Scoring of Knee Items used to generate the measure

1. Which treatment is most likely to provide relief from knee pain caused by osteoarthritis?

Surgery (Coded- 1)

Non-surgical treatments (coded =0)

Both are about the same (coded= 0)

2. After knee replacement surgery, about how many months does it take most people to get back to doing their usual activities?

Less than 2 months (coded= 0)

2 to 6 months (coded = 1)

7 to 12 months (coded= 0)

More than 12 months (coded= 0)

3.If 100 people have knee replacement surgery, about how many will have less knee pain after the surgery?

20 (coded= 0)

40 (coded= 0)

60 (coded= 0)

80 (coded = 1)

4.If 100 people have knee replacement surgery, about how many will have a serious complication within 3 months after surgery?

4 (Coded=1)

10 (coded= 0)

14 (coded= 0)

20 (coded= 0)

5. If 100 people have knee replacement surgery, about how many will need to have the same knee replaced again in less than 15 years?

More than half (coded= 0)

About half (coded= 0)

Less than half (coded =1)

Scoring of Preference Item for Knee:

6. Which treatment did you want to have to treat your knee osteoarthritis?

Surgery (coded=1)

Non-surgical treatments (coded= 0)

Not sure (coded= 0)

Scoring of Hip Items used to generate the measure:

1. Which treatment is most likely to provide relief from hip pain caused by osteoarthritis?

Surgery (Coded- 1)

Non-surgical treatments (coded =0)

Both are about the same (coded= 0)

2. After hip replacement surgery, about how many months does it take most people to get back to doing their usual activities?

Less than 2 months (coded= 0)

2 to 6 months (coded = 1)

7 to 12 months (coded= 0)

More than 12 months (coded= 0)

3. If 100 people have hip replacement surgery, about how many will have less hip pain after the surgery?

30 (coded= 0)

50 (coded= 0)

70 (coded= 0)

90 (coded = 1)

4. If 100 people have hip replacement surgery, about how many will have a serious complication within 3 months after surgery?

4 (Coded=1)

10 (coded= 0)

14 (coded= 0)

20 (coded= 0)

5. If 100 people have hip replacement surgery, about how many will need to have the same hip replaced again in less than 20 years?

More than half (coded= 0)

About half (coded= 0)

Less than half (coded =1)

Scoring of Preference Item for Hip:

6. Which treatment did you want to have to treat your hip osteoarthritis?

Surgery (coded=1)

Non-surgical treatments (coded= 0)

Not sure (coded= 0)

Knowledge: The responses are coded as indicated above. A total knowledge score is calculated by summing the five items, dividing by 5 and converting to percentage to get scores 0-100%. Missing answers are considered incorrect and scored as 0. Multiple responses (e.g. on paper survey) are considered incorrect and coded as 0. A total knowledge score is calculated for all surveys that have three or more knowledge items completed.

Preference item: Respondents who mark surgery are considered to indicate a clear preference for surgery. Respondents that mark either non surgical treatments or not sure, are not considered to have a clear preference for surgery. Missing responses are not counted. Multiple responses (e.g. on a paper survey) are considered "not sure" and coded as 0.

A positive assessment "yes" for decision quality requires a knowledge score of 60% or higher and a clear preference for surgery. Otherwise, decision quality is "no."

[Response Ends]

sp.25. Attach a copy of the instrument (e.g. survey, tool, questionnaire, scale) used as a data source for your measure, if available.

[Response Begins]

Copy of instrument is attached.

[Response Ends]

sp.26. Indicate the responder for your instrument.

[Response Begins]

Patient

[Response Ends]

sp.27. If measure testing is based on a sample, provide instructions for obtaining the sample and guidance on minimum sample size.

Examples of samples used for testing:

- *Testing may be conducted on a sample of the accountable entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.*
- *The sample should represent the variety of entities whose performance will be measured. The 2010 Measure Testing Task Force recognized that the samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, however, all types of entities whose performance will be measured should be included in reliability and validity testing.*
- *The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.*
- *When possible, units of measurement and patients within units should be randomly selected.*

[Response Begins]

Patients of a particular surgeon or at a particular clinical site (which could be a group of providers or a hospital or other surgical site) who had a primary knee or hip replacement surgery are identified from medical records, claims or in some other way. Sampling should allow time for immediate recovery, while attempting to survey shortly after the procedure, for example, by sampling eligible patients 1- 6 months after the procedure. Patients can be sampled sequentially, or a pool of such patients who had the procedure in a particular time period (e.g. in the last 3 months) can be created and sampled at a rate that produces the desired number of potential respondents.

The Decision Quality Instruments from which the measure is calculated can be used in a population-based sample, such as a sample of a population in a geographic area. Eligible respondents could be identified from claims (such as Medicare claims files) or based on patient self- reports of having had the procedures within some time frame.

The Decision Quality Instruments have also been used with patients shortly after a consult with an orthopedic surgeon to discuss joint replacement surgery but before surgery. However, there is often not consistent or detailed enough coding of visits to reliably identify patients after the visit but before having one of these procedures. As a result, at this time, the measure is proposed for use with patients who have had surgical treatment.

For knee and hip replacement surgery, rates of informed, patient-centered surgery varied from 37.9% to 59.5% across sites. A general population sample of patients who had knee and hip replacement surgery had rates of informed, patient-centered surgery of 18.8%. A sample size about 150 would be needed to detect differences in proportions of 10-15% for the measure (e.g. from 25% to 40%) with 80% power. This size difference is what we have observed between sites that do and do not make an effort to do shared decision making.

Proxy respondents are not permitted. The patients who receive the procedure should answer the survey questions. The survey is available in English and Spanish.

[Response Ends]

sp.28. Identify whether and how proxy responses are allowed.

[Response Begins]

Proxy respondents are not permitted. The patients who receive the procedure should answer the survey questions.

[Response Ends]

sp.29. Survey/Patient-reported data.

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

[Response Begins]

Eligible participants are identified by the clinician, clinical site or third party. The survey has been administered by mail, phone and online for patients to complete at home. A combination of mail, email and phone reminders are often needed to achieve adequate response rates. A third party vendor may also be used to administer the survey. We recommend that data not be accepted if response rates are lower than 50%. Calculate response rate as all those responding divided by all those invited to answer the survey questions (American Association for Public Opinion Research (AAPOR) response rate 4).

[Response Ends]

sp.30. Select only the data sources for which the measure is specified.

[Response Begins]

Instrument-Based Data

[Response Ends]

sp.31. Identify the specific data source or data collection instrument.

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

[Response Begins]

The measure is derived from responses to the Hip and Knee Decision Quality Instruments. These patient reported surveys have been administered by mail, phone, online, and through the health system patient portal platform.

A combination of mail, email, and phone reminders are often needed to achieve adequate response rates.

A third party vendor may also be used to administer the survey.

We have used these questions in English and Spanish.

[Response Ends]

sp.32. Provide the data collection instrument.

[Response Begins]

No data collection instrument provided

[Response Ends]

2ma.01. Indicate whether additional empirical reliability testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Reliability - Testing. Include information on all testing conducted (prior testing as well as any new testing).

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

Current Submission:

Updated testing information here.

Previous Submission:

Testing from the previous submission here.

[Response Begins]

Yes

[Response Ends]

2ma.02. Indicate whether additional empirical validity testing at the accountable entity level has been conducted. If yes, please provide results in the following section, Scientific Acceptability: Validity - Testing. Include information on all testing conducted (prior testing as well as any new testing).

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

Current Submission:

Updated testing information here.

Previous Submission:

Testing from the previous submission here.

[Response Begins]

Yes

[Response Ends]

2ma.03. For outcome, patient-reported outcome, resource use, cost, and some process measures, risk adjustment/stratification may be conducted. Did you perform a risk adjustment or stratification analysis?

[Response Begins]

No

[Response Ends]

2ma.04. For maintenance measures in which risk adjustment/stratification has been performed, indicate whether additional risk adjustment testing has been conducted since the most recent maintenance evaluation. This may include updates to the risk adjustment analysis with additional clinical, demographic, and social risk factors.

Please update the Scientific Acceptability: Validity - Other Threats to Validity section.

Note: This section must be updated even if social risk factors are not included in the risk adjustment strategy.

[Response Begins]

No additional risk adjustment analysis included

[Response Ends]

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

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

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

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

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

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

AND

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

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

- an evidence-based risk-adjustment strategy (e.g., risk models, risk stratification) is specified; is based on patient factors (including clinical and social risk factors) that influence the measured outcome and are present at start of care; 14,15 and has demonstrated adequate discrimination and calibration
- rationale/data support no risk adjustment/ stratification.

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

OR

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

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

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

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

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

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

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

Definitions

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

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

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

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

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

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

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

Current Submission:

Updated testing information here.

Previous (Year) Submission:

Testing from the previous submission here.

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

[Response Begins]

Instrument-Based Data

[Response Ends]

2a.02. If an existing dataset was used, identify the specific dataset.

The dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry).

[Response Begins]

Current submission:

Sample 3: Data collected as part of the orthopedic patient reported outcomes registry (n=3,470) patients surveyed as part of routine care shortly after hip or knee replacement surgery from four orthopedic practices affiliated with four hospitals (two academic medical centers and two community hospitals) that are part of a large health system.

Sample 4: DECIDE-OA data (n=559) patients with hip or knee osteoarthritis surveyed after orthopedic surgeon visit from three orthopedic practices affiliated with three sites (one academic medical center, one community hospital and one specialty hospital). Patients were participating in a randomized comparative effectiveness trial and all received decision aid as part of their care.

Sample 5: Patients (n=392) surveyed by mail within 6 months after hip or knee replacement surgery across a large health system with four orthopedic practices affiliated with four main hospitals (two academic medical centers and two community hospitals).

Previous submission:

Sample 1: A sample of 382 patients with hip and knee osteoarthritis were surveyed about one year after surgery or one year after discussing surgery with a surgeon. The respondents came from 3 different orthopedic groups in the Northeast, one of which was using decision aids and encouraging shared decision making for joint replacement surgery, a fourth group was general population sample who responded to a newspaper ad for the research study. A subset of respondents was sent the same survey 4-6 weeks later to examine retest reliability.

Sample 2: A sample of 127 patients who were part of a randomized controlled trial of knee and hip osteoarthritis patient decision aids were used to examine discriminant validity of the knowledge component of the measure. Participants were selected from an academic medical center in Canada.

[Response Ends]

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

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

[Response Begins]

Current submission:

Sample 3: 09-12-2018 through 05-21-2022

Sample 4: 05-2-2016 through 02-28-2018

Sample 5: 07-05-2018 through 12-07-2018

Previous submission:

2009-2010

[Response Ends]

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

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

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

Please do not select:

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

[Response Begins]

Clinician: Group/Practice

[Response Ends]

2a.05. List the measured entities included in the testing and analysis (by level of analysis and data source).

Identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample.

[Response Begins]

Current submission:

Sample 3: Participants were from four orthopedic groups affiliated with four hospitals within one large health system, two academic medical centers and two community hospitals.

Sample 4: Participants were from three orthopedic groups affiliated with three sites: one academic medical center, one community hospital and one specialty orthopedic hospital.

Sample 5: Participants were from four orthopedic groups affiliated with four hospitals within one large health system, two academic medical centers and two community hospitals.

Previous (2016) Submission:

Sample 1: Participants were selected from orthopedic groups affiliated with three academic medical centers in the Northeast and from the community. The community sample responded to an advertisement in a local newspaper.

Sample 2: Participants were selected from an orthopedic practice affiliated with an academic medical center in Canada that was running a randomized controlled trial of hip and knee osteoarthritis decision aids.

[Response Ends]

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

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

[Response Begins]

Current Submission:

Sample 3: The sample includes 3,470 patients from 53 surgeons across 4 sites. All patients who had primary knee or hip replacement surgery at the sites were assigned the surveys. The surveys were only available in English. Patients were on average 67.7 years old (SD 9.5), 58.2% were female and 53% had knee replacement surgery.

Sample 4: The sample included n=559 patients. Respondents needed to be at least 21 years old; read and speak English or Spanish; have a diagnosis of hip or knee osteoarthritis; and attend a visit with an orthopedic surgeon. Patients with recent hip fracture or aseptic necrosis, rheumatoid or psoriatic arthritis and recent prior joint replacement surgery were excluded. For these analyses, we limited to those respondents who underwent surgery. Respondents were on average 65 years old, 57% were female, 67% were diagnosed with knee osteoarthritis, and were predominantly White, non-Hispanic (89%). The sample is described in more detail in Sepucha et al 2019.

Sample 5: The sample includes 392 patients who had recently undergone hip or knee replacement surgery with 22 surgeons across 4 sites. An algorithm identified those who were eligible and removed those who were ineligible (Giardina et al 2020). Patients were on average 66 years old (SD 9years), 54% female, 94% White, non Hispanic, and 51% had hip replacement surgery (versus knee surgery). The sample is described in more detail in Valentine et al 2021.

Sepucha K, Bedair H, Yu L, Dorrwachter JM, Dwyer M, Talmo CT, Vo H, Freiberg AA. Decision Support Strategies for Hip and Knee Osteoarthritis: Less Is More: A Randomized Comparative Effectiveness Trial (DECIDE-OA Study). J Bone Joint Surg Am. 2019 Sep 18;101(18):1645-1653. doi: 10.2106/JBJS.19.00004. PMID: 31567801; PMCID: PMC6887636.

Giardina JC, Cha T, Atlas SJ, Barry MJ, Freiberg AA, Leavitt L, Marques F, Sepucha K. Validation of an electronic coding algorithm to identify the primary indication of orthopedic surgeries from administrative data. BMC Med Inform Decis Mak. 2020 Aug 12;20(1):187. doi: 10.1186/s12911-020-01175-1. PMID: 32787849; PMCID: PMC7425151.

Valentine KD, Cha T, Giardina JC, Marques F, Atlas SJ, Bedair H, Chen AF, Doorly T, Kang J, Leavitt L, Licurse A, O'Brien T, Sequist T, Sepucha K. Assessing the quality of shared decision making for elective orthopedic surgery across a large healthcare system: cross-sectional survey study. BMC Musculoskelet Disord. 2021 Nov 19;22(1):967. doi: 10.1186/s12891-021-04853-x. PMID: 34798866; PMCID: PMC8605511.

Previous (2016) Submission:

Sample 1: The full sample included n=382 (79% response rate to mailed survey) and a subset n=91 (83% response rate) completed the retest survey about 4 weeks after the initial survey. Respondents were aged 40 years and older with a

diagnosis of hip or knee osteoarthritis who either had total joint replacement or had discussed surgery with their physician (and chosen not to have TJR), within the past two years. Individuals with rheumatoid arthritis, psoriatic arthritis, osteonecrosis, partial knee replacement, revision surgery, or bilateral knee surgery were excluded.

Sample 2: The full sample included 127 respondents (92% response rate to the phone survey). Adult patients with osteoarthritis of the hip or knee who met the guidelines for referral to an orthopaedic surgeon for total joint replacement (TJR) and had access to a TV with a VCR or DVD player were recruited for participation. Patients with inflammatory arthritis; a previous total joint replacement; or who were deaf, blind, cognitively impaired, or had a language barrier were excluded. After signing a consent form, patients were randomized to receive either a patient decision aid on TJR or usual care. Both groups were instructed to review the information at home and complete the decision quality survey items. Approximately one week after recruitment, a research assistant telephoned participants to record the answers. The research assistant made an average of four calls to participants to complete the survey.

[Response Ends]

2a.07. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing.

[Response Begins]

Current submission:

Sample 3 was used for reliability. Samples 3, 4, and 5 were used for validity.

Previous submission:

Sample 1 was used for reliability. Samples 1 and 2 were used for validity.

[Response Ends]

2a.08. List the social risk factors that were available and analyzed.

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

[Response Begins]

Current submission:

Table 2a.08: Demographic characteristics of patient respondents for the recent samples

Characteristic	Sample 3 N = 3,470	Sample 4 N = 559	Sample 5 N=392
Gender: Female, n (%)	2019 (58%)	316 (57%)	212 (54%)
Age, mean (SD)	68 (9)	65 (9)	66 (9)
Race/Ethnicity: White, non Hispanic, n (%)	n/a	515 (94%)	369 (94%)
Education: ≥ College graduate, n (%)	n/a	347 (63%)	n/a
Joint: Hip n (%)	1631 (47%)	216 (39%)	200 (51%)

Characteristic	Sample 3 N = 3,470	Sample 4 N = 559	Sample 5 N=392
Quality of Life: EQ5D, mean (SD)	n/a	0.61 (0.19)	n/a

n/a=not asked; SD=standard deviation; EQ5D=EuroQol-5 Dimension is a measure of general quality of life, scores range from -0.11 – 1.0 with higher scores indicating higher quality.

Previous (2016) Submission:

Table 1: Demographic characteristics of patient respondents for Sample 1 and Sample 2.

*	**Sample 1**	**Sample 2**	**Sample 2**
Characteristic	All patients N=382	Hip/Knee Control N=66	Hip/Knee PtDA N=61
Gender: Male n (%)	169 (44)	27 (40.9)	25 (40.9)
Age mean (SD)	62.7 (9.6)	66.1 (9.49)	64.3(10.16)
Race/Ethnicity n (%)	*	*	*
White	359 (95.5)	Not asked	Not asked
Education n (%)	*	*	*
≥ College graduate	209 (56)	40 (60.6)	39 (63.9)
Some college	94 (25.2)	Not asked	Not asked
High school or less	68 (18.1)	26 (39.4)	22 (36.1)
Missing	9 (2.4)	0	0
Income n (%)	*	*	*
<$30,000	78 (20.5)	5 (7.6)#	7 (11.5)#
$30,000-60,000	70 (18.3)	21 (31.8)	18 (29.5)
$60,000-100,000	89 (23.3)	13 (19.7)	21 (34.4)
Over $100,000	93 (24.3)	22 (33.3)	12 (19.7)
Missing	52 (13.6)	5 (7.6)	3 (4.9)
Married/Committed relationship n (%)	255 (67.8)	42 (63.6)	38 (62.3)
Months since decision median (IQR)	11 (7, 15)	Considering decision	Considering decision
Had (or preferred) Surgery n (%)	235 (61) Had surgery	49 (74.2) Preferred surgery	39 (63.9) Preferred surgery
Joint (knee vs. hip): Knee n (%)	201 (53)	61 (94)	59 (97)
WOMAC Pain Score mean (SD)	5.6 (4.6)	10.7 (4.2)	11.2 (4.0)

PtDA=decision aid group; SD=standard deviation; N/A=not asked; FT=fulltime; IQR: interquartile range; # measured < $20,000; $$ measured from $20,000; WOMAC=Western Ontario McMasters University Arthiritis Index is a measure of disease specific pain

*Cell intentionally left blank

[Response Ends]

Note: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a.09 check patient or encounter-level data; in 2a.010 enter "see validity testing section of data elements"; and enter "N/A" for 2a.11 and 2a.12.

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

Choose one or both levels.

[Response Begins]

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

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

[Response Ends]

2a.10. For each level of reliability testing checked above, describe the method of reliability testing and what it tests.

Describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used.

[Response Begins]

Current Submission:

1. At the practice level, we divided data within each site to samples with a minimum size of 50. We then calculated the % with IPC within each sample. The reliability was calculated as variability from site divided by total variability. This is a valid measure of reliability similar to the traditional method of calculation intra-class or intra-rater correlation coefficient (in this case the rater is the site). [See for example, Fleiss J. The Design and Analysis of Clinical Experiments (Wiley Series in Probability and Statistics). Canada: Wiley and Sons, 1999.]

Previous (2016) Submission:

1. At the item level, we measured test-retest reliability of the knowledge and preference items from same individuals 4-6 weeks apart. For the knowledge score we examined the intraclass correlation coefficient (ICC) of the knowledge score at time 1 and time 2. The ICC compares the variability of different ratings of the same subject to the total variation across all ratings and all subjects. For the preference item, we examined the kappa between the response at time 1 and response at time 2. The kappa *statistic* measures agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, since κ takes into account the agreement occurring by chance.
2. At the practice level, we randomly split patients at the same clinical site into groups of 25 or larger and correlated the scores; i.e. how well score from one sample's reports correlated with another sample's reports for same decision for same provider group.
3. At the practice level, we also divided data within each site to samples with a minimum size of 25. We then calculated the % with IPC within each sample. The reliability was calculated as variability from site divided by total variability. This is a valid measure of reliability similar to the traditional method of calculation intra-class or intra-rater correlation coefficient (in this case the rater is the site). [See for example, Fleiss J. The Design and Analysis of Clinical Experiments (Wiley Series in Probability and Statistics). Canada: Wiley and Sons, 1999.]

[Response Ends]

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

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

[Response Begins]

Current Submission:

1. At the practice level, we divided data within each site to samples with a minimum size of 54 patients. In all, 72 patient groups were created from 3416 patient reports. We then calculated the % with IPC within each sample. The reliability was calculated as variability from site divided by total variability. At the practice level, we had 4 groups (site 1 had 16 samples, site 2 had 26, site 3 had 26 and site 4 had 4) and the reliability was 0.735

Previous (2016) Submission:

1. The test-retest reliability of the knowledge score was examined in sample 1 and found to be ICC=0.81 (95% CI 0.71 to 0.87). The test-retest reliability of the item assessing preferred treatment was (Kappa = 0.801).
2. At the practice level, the total sample size is 26 (site 1 has 1 combination, site 2 has 21 combinations, site 3 has 1 combination and site 4 has 3 combinations (sample 1 vs. 2, 2 vs. 3, 1 vs. 3)) and the results of the correlation analyses were 0.805.
3. At the practice level, we had 14 groups (site 1 had 2 samples, site 2 had 7, site 3 had 2 and site 4 had 3) and the reliability was 0.853.

[Response Ends]

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

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

[Response Begins]

Current submission:

With additional, larger samples, the reliability of the measure at the clinical practice level continues to be adequate.

Of note, the reliability estimate is slightly lower than the prior submission. We suspect that this difference is largely due to the randomization of individuals to groups. For example, if we randomize these same individuals 10 times, we find reliability ranges from 0.73 to 0.84, with a mean reliability value of 0.78 and 95% confidence interval (0.75, 0.80). If this same randomization was carried out on the prior data, we believe that the confidence intervals of the prior data and the current data would overlap, showing that these estimates are not considerably different.

Previous submission:

The test-retest reliability for the knowledge and preference items used to generate the measure is high. The reliability of the measure at the clinical practice level is also strong.

[Response Ends]

2b. Validity

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

[Response Begins]

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

Empirical validity testing

[Response Ends]

2b.02. For each level of testing checked above, describe the method of validity testing and what it tests.

Describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used.

[Response Begins]

Current Submission:

Three published studies provide additional evidence supporting validity of the measure (Sepucha et al 2018, Valentine et al 2021, Brodney et al 2019). Specifically, more evidence regarding:

1. Predictive validity of the overall IPC surgery measure.
    a. We hypothesized that patients who made IPC decisions would have more engagement in decisions (as measured by SDM Process scale), higher confidence (as measured by the SURE scale, a short form of the decisional conflict scale), higher satisfaction, and less regret. We used generalized linear and logistic regression models with the General Estimating Equations approach to account for clustering of patients within clinicians. Models adjusted for patient age, gender, education, joint, and baseline quality of life scores.
    b. We also tested hypotheses that IPC surgery is associated with better health outcomes using linear regression model with quality of life 6 months post surgery as the dependent variable and IPC, age , education, sex, treatment (surgery vs nonsurgery), joint (hip vs knee), site, baseline quality of life (SF-12 physical component score) as independent variables.

Sepucha KR, Atlas SJ, Chang Y, Freiberg A, Malchau H, Mangla M, Rubash H, Simmons LH, Cha T. Informed, Patient-Centered Decisions Associated with Better Health Outcomes in Orthopedics: Prospective Cohort Study. Med Decis Making. 2018 Nov;38(8):1018-1026. doi: 10.1177/0272989X18801308. PMID: 30403575.

Valentine KD, Cha T, Giardina JC, Marques F, Atlas SJ, Bedair H, Chen AF, Doorly T, Kang J, Leavitt L, Licurse A, O'Brien T, Sequist T, Sepucha K. Assessing the quality of shared decision making for elective orthopedic surgery across a large healthcare system: cross-sectional survey study. BMC Musculoskelet Disord. 2021 Nov 19;22(1):967. doi: 10.1186/s12891-021-04853-x. PMID: 34798866; PMCID: PMC8605511

Brodney S, Fowler FJ Jr, Barry MJ, Chang Y, Sepucha K. Comparison of Three Measures of Shared Decision Making: SDM Process_4, CollaboRATE, and SURE Scales. Med Decis Making. 2019 Aug;39(6):673-680. doi: 10.1177/0272989X19855951. Epub 2019 Jun 21. PMID: 31226911; PMCID: PMC6791732.

Previous submission:

The analyses replicate those published in Sepucha et al 2011 and Sepucha et al 2013 using the definition of the informed, patient centered hip and knee replacement surgery measure proposed here. The validity testing is done both at the individual component level (i.e. knowledge and preferred treatment) and at the measure level (i.e. informed, patient-centered (IPC) surgery).

1. A key feature of a knowledge test is that is can discriminate among those with different levels of knowledge and can detect clinically meaningful differences in knowledge resulting from interventions. As a result, we tested hypotheses that (a) providers would have higher knowledge scores than patients and that (b) patients who had seen a decision aid would have higher knowledge than the control group. Tested using two sample t-tests.
2. The validity of the item used to elicit preferred treatment was evaluated by seeing whether it discriminated patients' ratings of specific goals for pain relief, functional limitations and avoiding surgery. In other words, we examined whether patients who stated a clear preference for surgery rated the importance of relieving pain and improving function higher than those who were unsure or those who stated a preference for nonsurgical treatments. Further, we examined whether those who stated clear preference for surgery rated the importance

of avoiding surgery lower than those who were unsure or those who stated a preference for non surgical treatments. These hypotheses were tested using ANOVA with planned comparisons.

3. We tested the predictive validity of the overall IPC surgery measure. We hypothesized that patients who were informed and received treatments that matched their preferred treatment would have higher confidence (using a two sample t-test) and less regret (using a Chi squared test) than those who did not match.
4. We tested hypotheses that rates of IPC surgery are higher for patients who report more involvement in decision making process and are seen at a site that has formal decision support processes. We also tested hypotheses that IPC surgery is associated with better health outcomes. We first examined the following factors: age (<60 years vs_60 years), education (college or more vs other), sex,treatment (surgery vs nonsurgery), joint (hip vs knee), site, quality of life (SF-12 physical component score), and decision process score in univariate analyses using chi-square or t-tests, as appropriate. Then we developed a multivariable logistic regression model with high IPC surgery (yes/no) as the dependent variable and included all variables that were p<0.1 on univariate analyses as independent variables.

Sepucha KR, Stacey D, Clay CF, Chang Y, Cosenza C, Dervin G, Dorrwachter J, Feibelmann S, Katz JN, Kearing SA, Malchau H, Taljaard M, Tomek I, Tugwell P, Levin CA. Decision quality instrument for treatment of hip and knee osteoarthritis: a psychometric evaluation. BMC Musculoskelet Disord. 2011 Jul 5;12:149. doi: 10.1186/1471-2474-12-149. PMID: 21729315; PMCID: PMC3146909.

Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.

[Response Ends]

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

Examples may include correlations or t-test results.

[Response Begins]

Current Submission:

1.a. From Brodney et al 2019, we found that for hip and knee surgery decisions, IPC was significantly associated with higher shared decision making scores (mean SDM Process=2.3 non IPC vs. 2.7 IPC group, p<0.001) and higher decision confidence (SURE top score =63% non IPC vs. 92.3% IPC group, p<0.001). From Sepucha et al 2018, we found that controlling for age, sex, surgical status, education, and diagnosis (osteoarthritis v. spine), participants who made IPC decisions were more likely to be extremely satisfied with their pain (odds ratio [OR], 2.45; 95% CI, 1.45–4.15; P = 0.0008), were more likely to be very or extremely satisfied with their treatment (OR, 2.59; 95% CI, 1.59–4.22, P = 0.0001), and reported less regret (–5.63 points; 95% CI, –8.25 to –3.01; P \ 0.0001) than those who did not make IPC decisions.

1.b. Sepucha et al 2018 found that IPC was significantly associated with improvements in overall [0.05 points [SE 0.02] for EQ-5D, p=0.004) and disease-specific quality of life (4.22 points [SE 1.82] for knee p=0.02, and 4.46 points [SE 1.54] for hip, p=0.004). Sepucha et al 2022 found that IPC was related to overall (mean difference EQ-5D 0.04 points [0.02, 0.07], p<0.001) and disease specific quality of life (mean difference 4.9 points [1.5, 8.3], p=0.004) for knee but not hip patients.

Brodney S, Fowler FJ Jr, Barry MJ, Chang Y, Sepucha K. Comparison of Three Measures of Shared Decision Making: SDM Process_4, CollaboRATE, and SURE Scales. Med Decis Making. 2019 Aug;39(6):673-680. doi: 10.1177/0272989X19855951. Epub 2019 Jun 21. PMID: 31226911; PMCID: PMC6791732.

Sepucha KR, Atlas SJ, Chang Y, Freiberg A, Malchau H, Mangla M, Rubash H, Simmons LH, Cha T. Infor med, Patient-Centered Decisions Associated with Better Health Outcomes in Orthopedics: Prospective Cohort Study. Med Decis Making. 2018 Nov;38(8):1018-1026. doi: 10.1177/0272989X18801308. PMID: 30403575.

Sepucha KR, Vo H, Chang Y, Dorrwachter JM, Dwyer M, Freiberg AA, Talmo CT, Bedair H. Shared Decision-Making Is Associated with Better Outcomes in Patients with Knee But Not Hip Osteoarthritis: The DECIDE-OA Randomized Study. J Bone Joint Surg Am. 2022 Jan 5;104(1):62-69. doi: 10.2106/JBJS.21.00064. PMID: 34437308.

Previous submission:

1. We examined discriminant validity of the knowledge assessment by comparing scores of those who should have higher knowledge (e.g. scores of patients who had used a decision aid versus those who did not.) The mean knowledge scores discriminated between patients in decision aid group 67% (SD 21.2) compared to 51% (SD 24.9) in the usual care group ($p<0.001$). [Sepucha et al 2010]
2. To establish validity, we examined the extent to which patients' stated preference varied appropriately with specific goals. The table below provides evidence of the relationships in the predicted directions, supporting the validity of the single item as reflecting patients' preferred treatment.
3. Respondents had met the criteria for decision quality were more confident in their decision (9.09/10 vs. 7.78/10, $p<0.001$) and were significantly more likely to say they would do the same thing again (59.9%% vs. 26.4%%, $p<0.001$).
4. Replicating the multivariable logistic regression analyses from Sepucha 2013 [2] with the IPC surgery measure as proposed here, found the same results. None of the patient factors (age, sex, education) were significantly associated with IPC surgery. Controlling for treatment, IPC surgery was associated with more shared decision making and with the site that used decision aids. Further IPC surgery was significantly associated with higher quality of life as measured by the SF-12 Physical Component Score. The table below contains the results of these analyses.

Table: Patient stated treatment preference varied depending on their goals.

Question stem: On a scale of 1 to 10 where 1 is not at all important and 10 is extremely important,	Prefer surgery (N=218)	Unsure (N=26)	Prefer non surgical treatments	p (ANOVA)
…How important is it to relieve your knee pain?	9.50 (SD 1.19)	8.92 (SD 1.47)	8.43 (SD 2.42)	F=10.87, p<0.001
…How important is it not to be limited in what you can do because of your knee pain?	9.74 (SD 0.79)	9.38 (SD 1.33)	8.82 (SD 1.92)	F=12.37, p<0.001
…How important is it to you to avoid having surgery?	3.21 (SD 3.18)	5.50 (SD 2.92)	7.96 (SD 2.33)	F=71.65. p<0.001

Table: Results multivariate logistic regression with IPC surgery as dependent variable.

Variable	Odds Ratio	95% CI	p
Had Surgery	2.462	1.45, 4.17	0.001
Site (newspaper)	referent	*	0.16
Site 1	.896	.38, 2.10	.800
Site 2 (decision aid site)	2.275	1.22, 4.25	.010
Site 3	1.500	.69, 3.25	.305
Quality of life (SF-12 Physical component score)**	1.037	1.01, 1.06	.003
Shared decision making score**	1.012	1.00, 1.02	.015
College graduate	1.110	.67, 1.84	.686
Constant	.045	*	.000

*Cell intentionally left blank

**Odds ratio for a 10-point increase in scores.

[Response Ends]

2b.04. Provide your interpretation of the results in terms of demonstrating validity. (i.e., what do the results mean and what are the norms for the test conducted?)

[Response Begins]

Current submission:

The new data provide additional evidence supporting validity of the measure. The IPC surgery measure is significantly higher in practices with formal decision support than in those with limited use of decision support. Further, the IPC surgery measure demonstrated predictive validity and is associated with higher decision confidence, less regret, higher patient satisfaction, and better quality of life.

Previous submission:

The data provide evidence that the measure can discriminate among groups with different levels of knowledge (such as those who have viewed a decision aid or not), and the preference item can discriminate among patients with who place a different amount of importance on salient goals relating to treatment for osteoarthritis.

The IPC surgery measure is significantly higher in practices with formal decision support than in those without formal support. Further, the IPC surgery measure demonstrated predictive validity and is associated with higher confidence, less regret and better quality of life.

[Response Ends]

2b.05. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified.

Describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided in Importance to Measure and Report: Gap in Care/Disparities.

[Response Begins]

Current submission:

We have data from sites in the US that have formal shared decision making programs in routine care and sites that participated in a randomized trial comparing different types of decision support for patients that provide data on what is possible to achieve for the IPC measure. Previous submission identified magnitude of difference by comparing sites that did and did not have formal decision support available for patients.

Previous submission:

We compared the measure for practices that had implemented procedures to promote shared decision making and those who did not, including a general population sample. Multivariable logistic regression analyses were used to examine factors associated with rates of informed, patient-centered surgery.

A randomized controlled trial where the Hip and Knee Decision Quality Instruments were used also provides data on meaningful differences in rates of informed, patient centered surgery for patients who were or were not exposed to patient decision aids.

[Response Ends]

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

Examples may include number and percentage of entities with scores that were statistically significantly different fro m mean or some benchmark, different from expected; how was meaningful difference defined.

[Response Begins]

Current submission:

The evidence from one health system (sample 3), that has been focused on shared decision making and has decision aids available for patients (albeit with variable use), suggests that sites can achieve rates in the 70-80% range. The DECIDE-OA trial (sample 4) achieved high rates of IPC at the three sites (> 90%). The trial had research coordinators encouraging patients to review decision aids and reminding surgeons to engage patients in decisions. Achieving the same levels in routine care will likely be more challenging. As the prior submission stated, we still suggest a minimal meaningful difference in scores of 10%.

Previous submission:

There was considerable variation in rates of IPC surgery across sites, (31.8%, 50.0%, 56.0%, 64.7%) and in all cases, there was considerable room for improvement in rates. Compared to the general population referent group, the site that use d patient decision aids achieved significantly higher rates of IPC OR 2.275 (95% CI 1.22, 4.25) [Sepucha et al. 2013].

Two randomized controlled trials provide additional evidence for the potential magnitude of impact of decision aids on rates of IPC surgery. In the first, a randomized controlled trial with 142 patients found higher rates of IPC surgery in the intervention (patient decision aid) compared to control (pamphlet) group (56.4% intervention versus 25.0% control; $p < 0.001$) [Stacey et al. 2014]. In the second, a randomized controlled trial evaluating the same decision aids with 340 patients, rates of IPC surgery were also higher in the intervention (56.1%) compared to the control group (44.5%), relative risk (RR) 1.25; 95% CI 1.00-1.56, P = 0.050 [Stacey et al. 2016].

Based on the different randomized and non randomized studies, it is possible to see differences from 10%-30% in rates of IPC surgery across sites or groups of patients. From these data we suggest a minimal meaningful difference in scores of 10%.

Sepucha K, Feibelmann S, Chang Y, Clay CF, Kearing S, Tomek I, Yang TS, Katz JN. Factors associated with high decision quality for treatment of hip and knee osteoarthritis. J Am Coll Surg 2013 Oct;217(4):694-701. doi: 10.1016/j.jamcollsurg.2013.06.002. Epub 2013 Jul 25.

Stacey D(1), Hawker G, Dervin G, Tugwell P, Boland L, Pomey MP, O'Connor AM, Taljaard M. Decision aid for patients considering total knee arthroplasty with preference report for surgeons: a pilot randomized controlled trial. BMC Musculoskelet Disord. 2014 Feb 24;15:54. doi: 10.1186/1471-2474-15-54.

Stacey D(1), Taljaard M(2), Dervin G(3), Tugwell P(4), O'Connor AM(5), Pomey MP(6), Boland L(7), Beach S(8), Meltzer D(9), Hawker G(10). Impact of patient decision aids on appropriate and timely access to hip or knee arthroplasty for osteoarthritis: a randomized controlled trial. Osteoarthritis Cartilage. 2016 Jan;24(1):99-107. doi: 10.1016/j.joca.2015.07.024. Epub 2015 Aug

[Response Ends]

2b.07. Provide your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities.

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

[Response Begins]

Current submission:

As in the previous submission, there is considerable evidence that "usual care" results in fairly low rates of IPC surgery, suggesting considerable room for improvement. The evidence is pretty strong that this measure is a valid and reliable

assessment of the extent to which patients are well-informed and receive their preferred treatments. The evidence also supports the ability of sites that use patient decision aids or other shared decision making approaches to achieve consistently high scores.

Previous submission:

There is considerable evidence that "usual care" results in fairly low rates of IPC surgery, suggesting considerable room for improvement. The evidence is pretty strong that this measure is a valid and reliable assessment of the extent to which patients are well-informed and receive their preferred treatments. The evidence also supports the ability of existing tools (e.g. patient decision aids) to result in a meaningful improvement in the measure.

[Response Ends]

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

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

[Response Begins]

Current submission:

As before, we considered different approaches for handling missing data for the knowledge items. The first approach is to consider a missing answer as incorrect (with those responses coded as 0). The second approach is to impute the score of 1/k where k is the number of potential response options (essentially providing the points equivalent to guessing from the available multiple choice responses). We calculated the frequency of missing responses for each item in the knowledge assessment and then conducted sensitivity analyses to examine the impact on total knowledge scores.

Previous submission:

We considered different approaches for handling missing data for the knowledge items. The first approach is to consider a missing answer as incorrect (with those responses coded as 0). The second approach is to impute the score of 1/k where k is the number of potential response options (essentially providing the points equivalent to guessing from the available multiple choice responses). We calculated the frequency of missing responses for each item in the knowledge assessment and then conducted sensitivity analyses to examine the impact on total knowledge scores.

As described in section 2b3, 7/382 (1.8%) of respondents did not complete the preferred treatment item. We exclude respondents who do not complete that item and presented the results of those analyses in the earlier section.

[Response Ends]

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

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

[Response Begins]

Current submission:

Table 2b.09a. Missing knowledge and preference items from Samples 4 & 5.

item	Sample 4 N=568 missing n (%)	Sample 5 N=405 missing n(%)
Knowledge 1	7 (1.2 %)	9 (2%)
Knowledge 2	3 (0.5 %)	4 (1%)
Knowledge 3	10 (1.8 %)	15 (4%)
Knowledge 4	15 (2.6 %)	24 (6%)
Knowledge 5	10 (1.8 %)	18 (4%)
Preference	2 (0.4 %)	3 (1%)
Total missing IPC score	9 (1.6%)	13 (3%)

Table 2b.09b. Results from the different imputation methods for missing knowledge items for Sample 4

Number of knowledge questions answered	Frequency (%)	% score 60 or higher (missing as incorrect)	% score 60% or higher (missing with 1/k imputation)
0	1 (0.2%)	0 (0%)	0 (0%)
1	2 (0.4%)	0 (0%)	0 (0%)
2	4 (0.7%)	0 (0%)	0 (0%)
3	3 (0.5%)	2 (67%)	2 (67%)
4	14 (2.5%)	13 (93%)	13 (93%)
5	544 (96%)	517 (95%)	517 (95%)

Table 2b.09b. Results from the different imputation methods for missing knowledge items for Sample 5

Number of knowledge questions answered	Frequency (%)	% score 60 or higher (missing as incorrect)	% score 60% or higher (missing with 1/k imputation)
0	1 (0.2%)	0 (0%)	0 (0%)
1	1 (0.2%)	0 (0%)	0 (0%)
2	8 (2.0%)	0 (0%)	0 (0%)
3	11 (2.7%)	4 (36%)	4 (36%)
4	15 (3.7%)	11 (73%)	11 (73%)
5	369 (91.1%)	311 (84%)	311 (84%)

The Table shows overall frequency of missing data. Overall missing is small for both Sample 4: 9/568 (1.6%) and for Sample 5: 13/405 (3%). Patient characteristics (age, gender, race/ethnicity) did not vary significantly between those who were and were not missing data.

Previous submission:

The Table below shows the overall frequency of missing data for individual knowledge items. Twelve participants (3.1%) had 1 or 2 items missing and one respondent did not complete any items (0.3%). The knowledge scores are considerably lower for respondents with missing data; however the samples are very small.

Table: Missing responses and comparison of two approaches for handling missing data for the knowledge items used to generate the measure

Number of questions answered	Frequency (%)	% with Knowledge score 60% or higher (missing as incorrect)	% with Knowledge score 60% or higher (missing with 1/k imputation)
0	1 (0.3%)	0%	0%
1	0 (0%)	n/a	n/a
2	0 (0%)	n/a	n/a
3	2 (0.5%)	0%	0%
4	10 (2.6%)	30%	30%
5	368 (96.5%)	69.5%	69.5%

[Response Ends]

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

In other words, what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis was conducted, justify the selected approach for missing data.

[Response Begins]

Current submission:

We continue to find low missing data. As we found previously, missing data and the approach to treating missing data have a negligible impact on the rates of IPC surgery.

Previous submission:

Generally, missing data are low. Given the threshold for the indicator variable (correctly answering three or more items), the approach to missing data (either imputing 1/k or considering it incorrect) does not impact the % of respondents who meet that threshold. As a result, missing data and the approach to treating missing data have a negligible impact on the rates of IPC surgery.

[Response Ends]

Note: This item is directed to measures that are risk-adjusted (with or without social risk factors) OR to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eCQMs). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator). Comparability is not required when comparing

performance scores with and without social risk factors in the risk adjustment model. However, if comparability is not demonstrated for measures with more than one set of specifications/instructions, the different specifications (e.g., for medical records vs. claims) should be submitted as separate measures.

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

[Response Begins]

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

[Response Ends]

2b.12. Describe the method of testing conducted to compare performance scores for the same entities across the different data sources/specifications.

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

[Response Begins]

[Response Ends]

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

Examples may include correlation, and/or rank order.

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

[Response Begins]

Yes, the measure uses exclusions.

[Response Ends]

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

Describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used?

[Response Begins]

Current Submission:

As reported previously, respondents who skip 3 or more knowledge items or the preference item do not receive a total score.

Previous submission:

The IPC hip and knee replacement measure excludes surveys that have 3 or more knowledge responses missing or the preference item missing. To evaluate how missing data might affect validity we examined the frequency of *included* and *excluded* responses across patient characteristics including age, sex, education, and joint (hip or knee). To perform this analysis we created frequency distribution tables then performed a chi-square goodness of fit test. The chi-square tests the null hypothesis that there are no significant differences in the amount of included or excluded surveys between groups. If the test is significant to a p-value of 0.05 or less then we reject the null hypothesis and conclude there are significant differences between groups.

To evaluate the effect of exclusions across organizations, we examined the frequency of included and excluded responses for each site and tested for difference using a chi-square test.

We also calculated "expected" cell frequencies. The expected cell frequency represent the expected frequency of responses should the null hypothesis be true. This allows us to evaluate the departure from the expected number of excluded responses under the null hypothesis.

[Response Ends]

2b.17. Provide the statistical results from testing exclusions.

Include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores.

[Response Begins]

Current submission:

Table 2.b.17: Results from testing exclusions for missing data for Sample 4

*	No Missing	Missing IPC	p
N	559	9	*
Age <65 N (%)	266 (48 %)	3 (33 %)	0.51
Female N (%)	316 (57 %)	4 (44 %)	0.51
Practice 1 N (%)	108 (19 %)	1 (11 %)	0.74
Practice 2 N (%)	165 (30 %)	2 (22 %)	*
Practice 3 N (%)	286 (51 %)	6 (67 %)	*
Practice 4 N (%)	n/a	n/a	*
Joint: Hip N (%)	216 (39 %)	3 (33 %)	1
White, non-Hispanic	515 (94 %)	7 (78 %)	0.40
College Grad	347 (63 %)	3 (33 %)	0.16

*	No Missing	Missing IPC	p
High Literacy	392 (71 %)	4 (44 %)	0.14

**cells intentionally left blank*

Table 2.b.17: Results from testing exclusions for missing data for Sample 5

*	No Missing	Missing IPC	p
N	392	13	*
Age <65 N (%)	169 (43 %)	3 (23 %)	0.25
Female N (%)	212 (54 %)	11 (85 %)	0.04
Practice 1 N (%)	136 (35 %)	5 (38 %)	0.61
Practice 2 N (%)	130 (33 %)	4 (31 %)	*
Practice 3 N (%)	29 (7 %)	2 (15 %)	*
Practice 4 N (%)	97 (25 %)	2 (15 %)	*
Joint: Hip N (%)	200 (49 %)	9 (69 %)	0.26
White, non-Hispanic	369 (94 %)	12 (92 %)	0.55
College Grad	n/a	n/a	*
High Literacy	n/a	n/a	*

*Cell intentionally left blank

Previous submission:

We found very little missing data and as a result, there were very few exclusions. In sample 1, 2.1% or 8/382 respondents were excluded for not completing enough items.[1] Of those 8 exclusions, 7/8 did not complete the preference item and 1/8 did not complete at least 3 of the knowledge items.

Table: Included and excluded responses by characteristic with expected frequencies.

Variable (chi-square p-value)	Included (Expected) Column %	Excluded (Expected) Column %
Age (p=0.41)	*	*
Age >65	153 (151.5) 41.6%	1 (2.5) 16.7%
Age <65	215 (216.5) 58.4%	5 (3.5) 83.3%
Joint (p=0.49)	*	*

Variable (chi-square p-value)	Included (Expected) Column %	Excluded (Expected) Column %
Hip	176 (177.2) 47.1%	5 (3.8) 62.5%
Knee	198 (196.8) 52.9%	3 (4.2) 37.5%
Sex (p=0.74)	*	*
Male	165 (165.5) 44.1%	4 (3.5) 50%
Female	209 (208.8) 55.9%	4 (4.5) 50%
Education (p=0.17)	*	*
College or more	208 (206.2) 56.5%	1 (2.8) 20%
Less than college	160 (161.8) 43.5%	4 (2.2) 80%
Practice (p=0.82)	*	*
Practice 1	50 (49.9) 13.4%	1 (1.1) 12.5%
Practice 2	173 (174.3) 46.3%	5 (3.7) 62.5%
Practice 3	66 (65.6) 17.6%	1 (1.4) 12.5%
Practice 4	85 (84.2) 22.7%	1 (1.8) 12.5%

*Cell intentionally left blank

[Response Ends]

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

In other words, the value outweighs the burden of increased data collection and analysis. Note: If patient preference is an exclusion, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion.

[Response Begins]

Current submission:

As with the previous submission, we did not find significant or meaningful differences by site or patient characteristics. In Sample 5, gender was significant in one sample (suggesting females were more likely to have missing data) but the numbers were very small and we did not find similar result in Sample 4 (where females were less likely to have missing data). There is still limited power to detect differences for some characteristics (and due to the relatively small amount of excluded data).

Previous submission:

Overall, we found had few exclusions. We did not find any significant differences by site or by patient characteristics; however, with this sample size there was limited power to detect significant differences. Even if there were some statistically significant differences, the magnitude is likely to be very small so that the effect of those differences on results would be minimal and not likely sufficient to bias results.

[Response Ends]

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

[Response Begins]

No risk adjustment or stratification

[Response Ends]

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

[Response Begins]

[Response Ends]

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

[Response Begins]

Current submission:

We still do not recommend risk adjustment for this measure. Any patient who has one of these elective surgeries, should be able to answer the knowledge questions correctly and should have a clear preference for the procedure (to meet the standards of informed consent).

Previous submission:

We do not recommend risk adjustment for this measure. Any patient who has one of these elective surgeries, should be able to answer the knowledge questions correctly (to meet the standards of informed consent) and should have a clear preference for the procedure.

[Response Ends]

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

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

2b.27. Provide risk model discrimination statistics.

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

[Response Begins]

[Response Ends]

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

[Response Begins]

Not applicable.

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

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

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

[Response Begins]

[Response Ends]

Criterion 3. Feasibility

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

[Response Begins]

Other (Please describe)

[Other (Please describe) Please Explain]

Patient reported

[Response Ends]

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

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

[Response Begins]

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

[Response Ends]

3.03. If ALL the data elements needed to compute the performance measure score are not from electronic sources, specify a credible, near-term path to electronic capture, OR provide a rationale for using data elements not from electronic sources.

[Response Begins]

The patient-report surveys can be administered online to support electronic capture via patient reported outcomes registries or other online survey platforms. If administered via mail or paper, then it will require staff at sites to enter the patient data into an online database for analysis.

[Response Ends]

3.04. Describe any efforts to develop an eCQM.

[Response Begins]

At one health system, the items have been incorporated into the Patient-Reported Outcomes registry and are captured and scored as part of routine orthopedic care for patients undergoing surgery for hip, knee and spine conditions.

[Response Ends]

3.06. Describe difficulties (as a result of testing and/or operational use of the measure) regarding data collection, availability of data, missing data, timing and frequency of data collection, sampling, patient confidentiality, time and cost of data collection, other feasibility/implementation issues.

[Response Begins]

These data are from patient self-report. The administration of these questions has been conducted across multiple sites, in multiple modes (predominantly paper and online surveys). A large health system has incorporated the items into their patient-reported outcomes registry for orthopedics and the data is being collected as part of routine care in that system. Generally, patients find these surveys acceptable as indicated by good response rates and low missing data. However, whether administered as a stand-alone survey or as part of a patient-reported outcomes measure set, to obtain sufficiently high response rates often requires effort on the part of clinic staff (for example to remind patients to complete). Further, as mentioned in prior submission response below, it is easier to identify and survey patients who undergo surgery than those who pursue non-operative care.

Prior Submission: These questions have been extensively cognitively tested to ensure that they are consistently understood and that answers meaningfully describe patient experiences. We have used the questions proposed, and slight variations thereon, in a variety of survey designs: cross-section surveys of adults 40 and older, Medicare beneficiaries known to have had procedures based on claims, and clinical settings in which patients were identified by office staff or via medical records. The following observations have informed this proposal.

1. While we have included an "I am not sure" response with the knowledge items, particularly when used in the clinic at the time of initial decision making, when we have removed that option, the knowledge scores are higher as many patients do have a sense of the correct answer and will indicate it.
2. We can identify patients making decisions by asking them whether or not they had discussed an intervention, test or treatment. However, for cross-sections of adults or patients, the rates of any particular decision being made are too low to produce reliable data without very large samples.
3. We have surveyed patients in clinical settings before they had treatment. That is certainly the preferred way to measure informed, patient-centered surgery at a clinical site. However, it requires considerable integration into the clinic workflow and significant resources to get adequate response rates. It is easier to accomplish at sites that routinely assess patient-reported outcomes for all surgical patients (as the Decision Quality Instrument items can be included as part of the pre-operative assessment). It is also easier at sites that routinely use patient decision aids for their hip and knee osteoarthritis patients. In order to get comparable results across clinicians or clinical sites, we recommend sampling those patients who actually had the target intervention. In that way, patients can be reliably identified.
4. The hip and knee results are similar within sites, and as a result, we feel that it is reasonable to combine these two decisions in this measure.

[Response Ends]

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

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

Attach the fee schedule here, if applicable.

[Response Begins]

There are no fees for the measure or for the use of the Hip or Knee Decision Quality Instruments used to generate the measure, provided the surveys are used in accordance with the creative commons copyright license.

[Response Ends]

Criterion 4: Use and Usability

4a. Use

4a.01. Check all current uses. For each current use checked, please provide:

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

[Response Begins]

Payment Program

[Payment Program Please Explain]

Name of program and sponsor: Blue Cross Blue Shield of Massachusetts Alternative Quality Contract

URL: https://www.bluecrossma.org/aboutus/our-mission

Purpose: Blue Cross Blue Shield of Massachusetts (BCBSMA) is emphasizing measurement of decision quality and shared decision making to support their strategic goal of transitioning from legacy quality measures to measures which will better reflect care that is truly ethical, patient-centered, and high quality:

1. They have engaged Alternative Quality Contract provider groups in collecting pilot data using questionnaires based on the Decision Quality Instruments that form the basis for the IPC Hip/Knee Surgery measure.
2. They have proposed and been successful in adding these instruments to the Massachusetts Executive Office of Health and Human Services Aligned Measure Set for Accountable Care Organization contracts.
3. By the end of 2022, they are planning to collect data directly from their members using these instruments as a basis for confidential reporting to providers, with the goal of using these performance data as a basis for financial incentives.

Geographic area and number and percentage of accountable entities and patients included: Blue Cross Blue Shield of MA serves nearly 3 million members across MA and New England. The Alternative Quality Contracts with 13 provider groups in the region.

Level of measurement and setting: Clinician: Group/Practice

Professional Certification or Recognition Program

[Professional Certification or Recognition Program Please Explain]

Name of program and sponsor: The Alliance Quality Path Program

URL: https://the-alliance.org/quality-path/

Purpose: Quality Path Program sponsored by the Alliance specifies measurement of decision quality and shared decision making as part of their criteria for recognition. The purpose of the Quality Path program is to recognize providers and hospitals who are delivering high quality surgical care. The relevant section from the program detailing use of the measure is excerpted below and the entire program details can be found at the website link listed above.

Providers and practices are required to provide a description of the process for assessing the quality of shared decision making. This process needs to use the Decision Quality Instruments that form the basis for measure 2958. Ideally, for each procedure, practices will provide percentages, numerators, and denominators of patients participating in an assessment of shared decision making broken out by physician, practice, and by facility. Denominator is all patients receiving elective knee replacement or elective hip replacement. If the process has not been in place long enough to produce these numbers, this requirement may be waived until the six-month maintenance of designation process.

Geographic area and number and percentage of accountable entities and patients included: The Alliance is a cooperative of employers that includes more than 240 members who provide self-funded health benefits to more than

100,000 individuals. The network lets members choose from more than 80 hospitals, 13,500 total professional service providers, and 3,400 medical clinic sites in Wisconsin, Illinois, and Iowa.

Level of measurement and setting: Clinician: Group/Practice

Quality Improvement (Internal to the specific organization)

[Quality Improvement (Internal to the specific organization) Please Explain]

Name of program and sponsor: Shared Decision Making Program at Massachusetts General Brigham Health System

URL: https://www.massgeneralbrigham.org/en/about/newsroom/articles/shared-decision-making and https://mghdecisionsciences.org/tools-training/decision-quality-instruments/

Purpose: The Shared Decision Making Program sponsored in part by Mass General Brigham and Massachusetts Physician´s Organization has collaborated with the MGB Neurosurgery and Orthopedic Surgery Collaborative to incorporate the items IPC measure into the Patient Reported Outcomes Registry. All patients undergoing primary hip or knee replacement surgery are surveyed about 2-6 months after their procedure. Responses are summarized across surgeons and practices, used to identify high and low performing clinicians, and used to promote quality improvement initiatives in the departments. The initiative is also working to integrate patient decision aids into routine orthopedic care.

Geographic area and number and percentage of accountable entities and patients included: This project works with 6 hospitals and 158 surgeons, operating on about 5,800 patients annually within the Mass General Brigham system.

Level of measurement and setting: Clinician: Group/Practice

Use unknown

[Use unknown Please Explain]

The IPC measure 2958 is part of the aligned measure set that is available for use in payment programs in Massachusetts. It is not part of the core set, and as a result, it is not mandatory. Rather it is on the 'menu set' and available for use. At this time, we do not know whether any health systems, hospitals or other entities have selected to use this as part of their measures or whether any other insurers (aside from BCBS MA as described above) have incorporated the measures into their contracts.

Name of program and sponsor: Massachusetts Aligned Measure Set for Global Budget-Based Risk Contracts sponsored by Executive Office of Health and Human Services (EOHHS), the Massachusetts Health Policy Commission (HPC), and the Center for Health Information Analysis (CHIA)

URL: https://www.mass.gov/info-details/eohhs-quality-measure-alignment-taskforce

Purpose: The purpose is to recommend a set of measures to be used in global budget-based risk contracts for insurers and providers in Massachusetts. The Taskforce has developed an aligned measure set for voluntary adoption by private and public payers and by providers in global budget-based risk contracts. By doing so, the Taskforce strives to advance progress on state health priorities and reduce use of measures that don't add value. Contracts between payers (commercial and Medicaid) and provider organization where budgets for health care spending are set either prospectively or retrospectively, according to a prospectively known formula, for a comprehensive set of services for a broadly defined population, and for which there is a financial incentive for achieving a budget. The contract includes incentives based on a provider organization's performance on a set of measures of health care quality or there is a standalone quality incentive applied to the same patient population.

Geographic area and number and percentage of accountable entities and patients included: All commercial and Medicaid contracts in the state of Massachusetts.

Level of measurement and setting: Clinician: Group/Practice

[Response Ends]

4a.02. Check all planned uses.

[Response Begins]

Measure Currently in Use

[Response Ends]

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

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

[Response Begins]

The measure is currently being used in both a payment program and certification program as described above.

[Response Ends]

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

A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified timeframes. A plan for accountability applications addresses mechanisms for data aggregation and reporting.

[Response Begins]

The measure is currently being used in both a payment program and certification program as described above.

[Response Ends]

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

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

[Response Begins]

The measure developer and team at MGH have supported the administration of the measure, scoring and interpretation of results for the quality improvement initiatives. The team at MGH has created user guides that summarize the psychometrics of the measure, highlight issues regarding implementation and then also clarify scoring. The user guide is freely available from the MGH Health Decision Sciences website: https://mghdecisionsciences.org/tools-training/decision-quality-instruments/. Further, the MGH team is working with BCBS MA and their clients to refine sampling plans, confirm item wording and instructions, and will be available to provide assistance with interpretation of results.

[Response Ends]

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

[Response Begins]

For the quality improvement project, we are tracking response rates quarterly and providing feedback to the site champions quarterly and to the departments about twice a year. We have hosted several sessions describing the measures, interpreting results, and then also providing information on interventions (e.g. patient decision aids) that are available to surgeons to help increase scores.

The BCBS MA payment program is being led by the BCBS team in conjunction with their clients. While we are advising on that program, we are not directly involved in collecting or analyzing the data.

[Response Ends]

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

[Response Begins]

The administrative and clinical leaders in orthopedics found the short measure acceptable and relatively easy to incorporate into the PROMs registry. Patients have not complained about undue burden due to these 6-items being added. The existing online patient reported outcomes platform supported the routine collection. Further, adding the items to one of the existing time frames (as opposed to creating a new assessment) was critical to get buy-in. As a result, we are only surveying patients who underwent surgery at this time, as there is not consistent follow-up with PROMs for non-surgical patients. We meet quarterly with MGB leadership and also meet with the PROMs team to track feedback and identify any issues.

[Response Ends]

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

[Response Begins]

We have feedback from patients who have participated in research studies using these measures. As documented earlier, the measures are highly acceptable to patients with very little missing data. We have heard that patients are interested in the correct answers to the knowledge items and, when possible, we make those available after the assessment is completed. When we have shared results with the surgeons, we have had generally positive feedback. They often want to see the item-level responses to understand knowledge gaps or areas where patients have misperceptions that may be driving the scores and/or differences in the scores. The individual knowledge item results will identify areas where patients consistently have inaccurate understanding about options, benefits and harms. Occasionally, surgeons have challenged whether a particular knowledge answer is "correct." We are able to share the annotated evidence-base used to support the correct and incorrect responses. If they have new evidence, then we will consider changing the items and/or responses to reflect updated evidence. This open and transparent process often leads to them accepting the items and results.

[Response Ends]

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

[Response Begins]

The other users generally have positive reactions to the survey, they are happy that the survey is short and appreciate the simple scoring. Occasionally, we will receive questions regarding item wording and whether it is acceptable to make small changes to the items or instructions (for example, change 'health care providers' to 'surgeons'). We will consider these on case-by-case basis and review the context in which the items are being used before approving any wording changes.

[Response Ends]

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

[Response Begins]

We have used the feedback to update the user guide where we provide advice to users on how to best set up the survey to ensure high response rates and high quality data. The main advice has been to incorporate the survey items into existing registries or patient survey platforms supported by electronic medical records. In addition, we advise groups to be prepared to share the correct answers to the knowledge items after the surveys have been completed.

[Response Ends]

4b. Usability

4b.01. You may refer to data provided in Importance to Measure and Report: Gap in Care/Disparities, but do not repeat here. Discuss any progress on improvement (trends in performance results, number and percentage of people receiving high-quality healthcare; Geographic area and number and percentage of accountable entities and patients included). If no improvement was demonstrated, provide an explanation. If not in use for performance improvement at the time of initial endorsement, provide a credible rationale that describes how the performance results could be used to further the goal of high-quality, efficient healthcare for individuals or populations.

[Response Begins]

The measure was not in use for performance improvement at the time of initial submission. As described earlier, several published studies using the IPC measure have found patients provided with decision support interventions have significantly higher rates compared to usual care (Jayakumar et al. 2021, Stacey et al. 2014, Stacey et al. 2015). More recent studies (data presented in earlier sections of the submission) have also shown that when patients receive decision aid as part of routine care, that scores can be quite high (91%-95%); where as in practices with few patients receiving decision aids, scores are much lower (72-80%).

Patient-centered care is a core component of high quality health care. Definitions of patient-centered care emphasize the importance of informing and involving patients in medical decisions and ensuring that patients' goals and preferences are respected. This is particularly important in cases of elective surgery, where there is no definitive clinical need, and the use of surgery must be determined by informed patient preference. This measure provides a means to assess the extent to which patients who had elective surgery were well informed and had a clear preference for surgery.

References:

1. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, Aksan N, Rathouz PJ, Bozic KJ. Comparison of an Artificial Intelligence-Enabled Patient Decision Aid vs Educational Material on Decision Quality, Shared Decision-Making, Patient Experience, and Functional Outcomes in Adults With Knee Osteoarthritis: A Randomized Clinical Trial. JAMA Netw Open. 2021 Feb 1;4(2):e2037107. doi: 10.1001/jamanetworkopen.2020.37107. PMID: 33599773.
2. Stacey D, Hawker G, Dervin G, Tugwell P, Boland L, Pomey MP, O'Connor AM, Taljaard M. Decision aid for patients considering total knee arthroplasty with preference report for surgeons: a pilot randomized controlled trial. BMC Musculoskelet Disord. 2014 Feb 24;15:54. doi: 10.1186/1471-2474-15-54. PMID: 24564877; PMCID: PMC3937455.
3. Stacey D, Taljaard M, Dervin G, Tugwell P, O'Connor AM, Pomey MP, Boland L, Beach S, Meltzer D, Hawker G. Impact of patient decision aids on appropriate and timely access to hip or knee arthroplasty for osteoarthritis: a randomized controlled trial. Osteoarthritis Cartilage. 2016 Jan;24(1):99-107. doi: 10.1016/j.joca.2015.07.024. PMID: 26254238.

[Response Ends]

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

[Response Begins]

We have not encountered unintended negative consequences during use of the measure. We have heard from patients who are interested in seeing the correct answers to the knowledge items, as some report they did not learn this from their health care team. Since the implementation of the measure, we have also had several surgical colleagues approach us with interest in using the data to evaluate the decision-making process, answer important research questions and use the data to design improvements. For example, one of the surgeons involved in the MGB quality improvement work, who is also a medical director with CRICO, the malpractice insurer for Harvard physicians, plans to use the measure data in a larger quality improvement project to redesign the informed consent process and is exploring the use of this SDM Process measure (#2962) and the Informed Patient Centered Hip/Knee Replacement (#2958) measure as part of that work. Another surgeon is interested in looking at the data throughout the COVID pandemic to examine whether communication of information and patient comprehension was different for virtual visits compared to in-person visits, a project that is possible to do because we had incorporated the assessment into the PROMs system.

[Response Ends]

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

[Response Begins]

No other unexpected benefits aside from the experiences mentioned in 4b.02.

[Response Ends]

Criterion 5: Related and Competing Measures

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

(Can search and select measures.)

[Response Begins]

[Response Ends]

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

(Can search and select measures.)

[Response Begins]

[Response Ends]

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

[Response Begins]

Not applicable.

[Response Ends]

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

[Response Begins]

No

[Response Ends]

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

[Response Begins]

Not applicable.

[Response Ends]

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

Provide analyses when possible.

[Response Begins]

Not applicable.

[Response Ends]