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* and *Preliminary Analysis* 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 #: 2962

Corresponding Measures:

Measure Title: Shared Decision Making Process

Measure Steward: Massachusetts General Hospital

sp.02. Brief Description of Measure: This measure assesses the extent to which health care providers actually involve patients in a decision-making process when there is more than one reasonable option. While we believe that the survey will work for patients who have undergone any elective surgical procedure, we have proposed a limited set of surgeries based on existing data for these conditions. This measure focuses on patients who have undergone one of 7 common, important surgical procedures: total hip or knee replacement for osteoarthritis, lower back surgery for lumbar spinal stenosis or herniated disc, radical prostatectomy for prostate cancer, mastectomy for early stage breast cancer or percutaneous coronary intervention (PCI) for stable angina. Patients answer four questions (scored 0 to 4) about their interactions with providers about the decision to have the procedure, and the measure of the extent to which a provider or provider group is practicing shared decision making for a particular procedure is the average score from their responding patients who had the procedure.

1b.01. Developer Rationale:

We have collected a great deal of data from surveys of patients who have made decisions drawn from the general population and from clinical sites documenting that for many decisions, patients routinely do not perceive that they discuss the cons of proposed interventions, are not told about alternatives and are not asked to share their treatment preferences as part of the decision. Consistently, their levels of knowledge of information relevant to the decisions they are making are low. We then have evidence that when clinicians commit to shared decision making, by routinely providing decision aids for example, the scores of patients with respect to knowledge and the decision making process are higher. We believe the use of the Shared Decision Making Process Score and appropriate measures of patient knowledge can be catalysts to routinely informing and involving patients in important medical decisions, which in turn will increase the likelihood that patients will get the care they want and that is consistent with their goals and concerns (Stacey et al 2020).

Stacey D, Légaré F, Boland L, Lewis KB, Loiselle MC, Hoefel L, Garvelink M, O'Connor A. 20th Anniversary Ottawa Decision Support Framework: Part 3 Overview of Systematic Reviews and Updated Framework. Med Decis Making. 2020 Apr;40(3):379-398. doi: 10.1177/0272989X20911870. PMID: 32428429.

sp.12. Numerator Statement: Patient answers to four questions about whether or not 4 essential elements of shared decision making (laying out options, discussing the reasons to have the intervention, discussing reasons not to have the intervention, and asking for patient input) are scored and summed. A group/practice score is the average of their patient scores.

sp.14. Denominator Statement:

While we believe that the survey will work for patients who have undergone any elective surgical procedure, we have proposed a limited set of surgeries based on existing data for these conditions.

All responding patients who have undergone one of the following 7 surgical procedures: back surgery for a herniated disc; back surgery for spinal stenosis; knee replacement for osteoarthritis of the knee; hip replacement for osteoarthritis of the hip; radical prostatectomy for prostate cancer; percutaneous coronary intervention (PCI) for stable angina, and mastectomy for early stage breast cancer.

sp.16. Denominator Exclusions:

For back, hip, knee, and prostate surgery patients, there are no exclusions as long as the surgery is for the designated condition (for example, hip replacement for osteoarthritis not for hip fracture).

For PCI, we are focused on patients who are treated for stable coronary artery disease. As such, those who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).

For mastectomy, we are focused on females having mastectomy as the primary surgical treatment for breast cancer. Patients who had had a prior lumpectomy for breast cancer in the same breast, patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies), and males with breast cancer are excluded.

Respondents who are missing one or more responses to the SDM Process measure do not receive a total score and thus, are excluded.

Measure Type: Outcome: PRO-PM

sp.28. Data Source:

Instrument-Based Data

sp.07. Level of Analysis:

Clinician: Group/Practice

IF Endorsement Maintenance – Original Endorsement Date: 10/25/2016

Most Recent Endorsement Date: 10/25/2016

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) measure at the clinician group/practice level assesses the extent to which health care providers actually involve patients in a decision-making process when there is more than one reasonable option.
- The developer does not provide a logic model graphic, the developer describes the process of shared decision making (SDM) wherein clinicians meaningfully engage patients in medical decisions. The developer goes on to note that SDM involves helping patients recognize that there is a choice to be made, ensuring patients understand the pros and cons of the options and incorporating what matters most to patients into the final choice.
- The developer states that the goal of shared decision making is to improve decision quality, ensuring that decisions are well informed and reflect patient goals, concerns and preferences as has been associated with lower decisional conflict as well as less decision regret.

Summary of prior review in 2017

- During the 2017 measure evaluation meeting, the Person- and Family-Centered Care Standing Committee agreed that this PRO-PM demonstrated the value of the shared decision-making approach and the 4 items within the questionnaire are based on the 3 essential concepts it was designed to address.
- The developer noted that this measure works best when applied to a specific kind of decision (e.g. decision to have surgery for herniated disc).

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 2021 meta-analysis of the SDM Process scale for surgical decisions where researchers found that SDM Process scores were associated with higher decision quality, less decisional conflict, and lower decision regret.
- The developer also cites a study that focused on hip and knee replacement and spine surgery decision, where researchers found that SDM Process scores were related to less regret and higher patient satisfaction.

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 provides gap data from three sources: sample 4, sample 5, and sample 6.
    - Sample 4 consisted of the randomized DECIDE-OA study which surveyed 568 patients with hip or knee osteoarthritis 6 months after a orthopedic surgeon visit at one academic medical center, one community hospital and one specialty hospital from 04-19-2016 - 2-28-2018.
        - Patients were participating in a randomized comparative effectiveness trial and all received decision aid as part of their care.
        - Overall mean scores across clinical topics were 2.6 (hip) and 2.4 (knee).
        - Overall standard deviations across clinical topics were 1.1 (hip and knee).
    - Sample 5 included 646 patients surveyed by mail within 6 months after hip or knee replacement surgery across two academic medical centers and two community hospitals from 04-19-2016 - 2-28-2018.
        - Overall mean scores across clinical topics were 2.6 (hip), 2.5 (knee), 3.2 (herniated disc), and 3.1 (spinal stenosis).
        - Overall standard deviations across clinical topics were 1.2 (hip) and 1.1 (knee, herniated disc, spinal stenosis).
    - Sample 6 includes data collected an orthopedic patient reported outcomes registry where 5,330 patients were surveyed as part of routine care shortly after hip or knee replacement surgery or back surgery across a health system with two academic medical centers and two community hospitals from 04-19-2016 - 2-28-2018.
        - Overall mean scores across clinical topics were 2.9 (hip), 2.8 (knee), 3.3 (herniated disc), and 3.1 (spinal stenosis).
        - Overall standard deviations across clinical topics were 1.0 (knee, hip, and spinal stenosis) and 0.9 (herniated disc).
    - The developer concludes that the data shows fairly good scores for the orthopedic topics from sites that have been making an effort to engage patients in shared decision making as part of routine care.
    - The developer also states that there is some room for improvement in scores, particularly for hip and knee replacement surgery decisions.

Disparities

- The developer provides the following mean scores regarding disparities:
- Hip Replacement Surgery
    - Age in Sample 4 (<65 2.5 v. 65+ 2.4, p=0.34), Sample 5 (<65 2.5 v. 65+ 2.5, p=0.86), Sample 6 (<65 3.0 v. 65+ 2.7, p=0.00).
    - Gender in Sample 4 (Female 2.4 v. Male 2.5, p=0.74), Sample 5 (Female 2.5 v. Male 2.5, p=0.90) and Sample 6 (Female 2.7 v. Male 2.95, p=0.00).
    - Race Sample 4 (White, non-Hispanic 2.4 v. Other Race/ethnicity 2.5, p=0.89), Sample 5 (White, non-Hispanic 2.5 v. Other Race/ethnicity 2.5, p=0.97).
    - Education in sample 4 (College degree or more 2.5 v. Less than college degree 2.4, p=0.54).
- Knee Replacement Surgery
    - Age in Sample 4 (<65 2.7 v. 65+ 2.5, p=0.20), Sample 5 (<65 2.6 v. 65+ 2.6, p=0.73), Sample 6 (<65 3.0 v. 65+ 2.8, p=0.004).
    - Gender in Sample 4 (Female 2.55 v. Male 2.6, p=0.58), Sample 5 (Female 2.6 v. Male 2.6, p=0.78) and Sample 6 (Female 2.8 v. Male 2.8, p=0.001).

    - Race Sample 4 (White, non-Hispanic 2.55 v. Other Race/ethnicity 3.1, p=0.03), Sample 5 (White, non-Hispanic 2.6 v. Other Race/ethnicity 2.9, p=0.42
    - Education in sample 4 (College degree or more 2.6 v. Less than college degree 2.6, p=0.68).
- Herniated Disc Surgery
    - Age in Sample 5 (<65 3.25v. 65+ 2.9, p=0.33), Sample 6 (<65 3.4 v. 65+ 3.2, p=0.00).
    - Gender in Sample 5 (Female 3.2 v. Male 3.2, p=0.83) and Sample 6 (Female 3.2 v. Male 3.4, p=0.02).
    - Race Sample 5 (White, non-Hispanic 3.2 v. Other Race/ethnicity 3.75, p=0.29).
- Spinal stenosis Surgery
    - Age in Sample 5 (<65 3.2 v. 65+ 3.0, p=0.44), Sample 6 (<65 3.2 v. 65+ 3.1, p=0.01).
    - Gender in Sample 4 (Female 3.0 v. Male 3.1, p=0.59), Sample 5 (Female 3.0 v. Male 3.2, p=0.002).
    - Race Sample 5 (White, non-Hispanic 3.1 v. Other Race/ethnicity 2.9, p=0.59).
- The developer reports that younger respondents and males appear to have slightly higher scores, though most results are neither statistically nor clinically significant.
- The developer also reports that while sample 6 is large and most characteristics do reach statistical significance, the magnitude of the differences in SDM Process scores is small.
- The developer concludes that the findings do not support disparities for education or race/ethnicity, but also notes that the samples are also small for race/ethnicity.

Questions for the Standing Committee:

- *Is there a gap in care that warrants a national performance measure?*
- *If limited disparities information is provided, are you aware of evidence that disparities exist in this area of healthcare?*

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-0; M-8; L-0; I-2.
- The SMP passed on Validity with a score of: H-3; M-4; L-1; I-2

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.

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

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 developers divided patients from the same site making the same decisions into random groups and correlated their process scores.
    - The developer implemented a minimum sample size of 58, which results in 76 patient groups from 5,294 patient reports.
    - The developers reported an average reliability of 0.69 (95 percent CI = [0.685, 0.69])
    - The developers also reported an ICC by dividing the between site variance by the total variance resulting in an ICC of 0.96.
- Reliability testing was conducted at the patient/encounter level:
    - In the 2016 submission, the developer noted that Cronbach alpha may not be an appropriate measure of reliability due to the nature of the measure; however, they calculated the alphas for some decisions, noting that they are often in the 0.5–0.7 range.
    - The developer noted that the short-term, test-retest data on some variations of the measure obtained ICC values ranging from 0.7–0.8.
    - The developer also conducted tests of agreement, noting that in two tests of whether patient reports of their interactions align with the coding of tape recordings of the interactions, the level of agreement was high, although patient's ratings tended to be a bit higher than the observers'.
    - Additionally, in a different test of agreement, women's interactions with physicians about primary treatment for breast cancer were tape recorded. Coding of the interactions was related to patient reports using the questions in the Process Score. The developer notes that because the clinically reasonable options were known, questions were asked separately for a discussion of the pros and cons of both reasonable options. For this test, Kappas for dichotomous variables and product moment correlations for the multi-category items were reported.
        - Overall scores: correlations were 0.50 ($p<0.001$) for adjuvant therapy and 0.38 ($p=0.004$) for surgery decisions
        - Individual items:
            - Values were higher for whether options were presented (0.64–0.71) and how much the reasons for each option were discussed (0.64–0.75)
            - Values were lower for how much the cons were discussed (0.16–0.46) and whether the patient's input was sought (0.14–0.32)
    - Lastly, the developer noted that the previous average reliability at the clinician level with a minimum sample size of 25 was 0.61.

SMP Summary:

- One SMP member sought clarification regarding the specifications, specifically, how the measure scores are calculated when multiple types of surgeries (hip, knee, back) are involved. For example, is the intention to calculate the measure by condition?
- One SMP member questioned whether the patient/encounter level reliability testing was conducted on patients undergoing PCI, as this would be required if they are included in the denominator.
- There were a number of SMP members who were concerned that the accountable entity level reliability testing did not demonstrate adequate reliability for all surgery types (namely, prostate surgery, PCI, and mastectomy).

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 patient/encounter level:
    - In the current submission, the developer provides evidence from three published studies, which depict the relationship of this measure in the predicted direction with other decision-making outcomes (e.g., higher confidence; satisfaction; less regret; and higher rates of informed, patient-centered surgery).
        - In a 2021 paper, an effect size was calculated using a model of inverse variance methods and random effects. The heterogeneity of the effects was calculated using the DerSimonian–Lair estimator of between-study variance.
            - The developer attests that the paper's results showed that the SDM process scores were related to higher decision confidence (effect size = 0.57, $p<0.001$); lower decision regret (effect size = -0.34, $p<0.001$); and higher rates of informed, patient-centered decisions (effect size = 0.18, $p=0.03$).
        - In a 2019 paper, the developers used generalized linear and logistic regression models with the General Estimating Equations approach to account for the clustering of patients within surgeons.

- The developer states that models were adjusted for patient characteristics, such as age, gender, education, joint, and baseline quality of life scores.
- The developer attests that the 2019 paper's results showed that SDM process scores were higher among patients who reported no regret (2.5 [1.2] no regret versus 2.3 [1.2] regret, p<0.001 for hip and knee surgery); higher among patients who reported high satisfaction ([2.3 (1.2) not satisfied vs. 2.5 satisfied (1.2), p<0.001) for hip and knee surgery] and [(2.1 (1.4) not satisfied versus 2.6 (1.2) satisfied , p<0.001 for back surgery]); and were significantly higher for patients who made informed, patient-centered decisions compared to those who did not (2.7 versus 2.3, p<0.001 for hip and knee surgery and 3.2 [0.9] versus 2.0 [1.3], p<0.001 for back surgery).

- In another 2021 paper, a generalized linear and logistic regression model with the General Estimating Equations approach was used to account for clustering of patients within surgeons in a cross-sectional sample to identify relationships between the scale and health outcomes.
  - The developer attests that this paper's results showed that higher SDM process scores were associated with larger improvements from pre- to post-surgery in mental (b=0.16, p=0.02) and physical health (b=0.25, p=0.02) outcomes for patients who had total joint replacement of the hip or knee but not patients who had spine surgery (all p's greater than 0.26).

- Validity testing was conducted at the accountable-entity level:
  - The developer cites two studies at the site level, and one study summarized performance at the group/practice level. The developers noted they tested whether clinical practices that implemented shared decision making had higher SDM scores than sites practicing usual care.
  - The developers used t-tests to compare mean SDM scores from different settings using a Welch's correlation when needed. They also calculated Cohen's d effect sizes for all comparisons. The developer notes that a 0.2 effect size would indicate a small effect, 0.5 indicates a medium effect, and 0.8 indicates a large effect.
    - The developer notes that for osteoarthritis of the knee and hip, patients in the practices where decision aids were used reported significantly better decision processes (2.9 versus 2.5, P <0.001, d=0.49 and 2.9 versus 2.1, P <0.001, d= 0.84, respectively).
    - The developer notes that the difference in the SDM Process Scores for spine practices that did and did not use decision support (3.0 versus 2.75, P=0.12, d=0.22) was in the expected direction but was not large enough to reach statistical significance.
    - Lastly, the developer notes that with regard to breast cancer practices, the practice that had formal decision support had significantly better scores than cancer practices without any decision support interventions (2.7 versus 2.3, P< 0.05, d=0.47).
  - The developers also presented content validity from the 2021 paper, which noted that patients were unable to adequately describe shared decision making and their general desire to rate their clinicians highly, which proved to be problematic as the patients lacked a frame of reference for evaluating decision making. The developer notes that the findings resulted in the SDM process survey's focus on clinical decision and on the report of events or behaviors.
  - In the previous submission, the developer compared the aggregate SDM Process Score from patients treated at clinical sites that have committed to shared decision making with reports of national cross-sections of patients from the TRENDS survey who made the same decisions and compared the mean SDM scores for four breast cancer clinical sites where three used usual care

and one used decision aids. They also compared the mean SDM scores for hip and knee replacement sites that used usual care versus decision aids. Lastly, they compared SDM scores for a clinical site for patients who discussed treatment benign prostatic hyperplasia (BHP) for before the use of decision aids and after the use of decision aids.

- The developer states that the results indicate that clinical sites who commit to improved decision making attain average scores from their patients that are higher than the average.

Exclusions

- The developer notes that they do not send surveys to patients with exclusion codes, and as a result, they do not have data to test relating to those codes.
- The developers additionally note that they recommend excluding those who miss one or more of the SDM process items.
- When examining the impact of this exclusion, the developer found negligible impact on the performance scores due to the small number of those excluded.

Risk Adjustment

- The measure is not risk-adjusted or stratified.

Meaningful Differences

- To determine meaningful differences, the developer examined the differences between site-level scores with multivariable linear regression analyses with Generalized Estimating Equations to correct for correlated error due to patients being nested within surgeons. The developer noted that several studies show that the SDM Process survey has effect sizes ranging from 0.39SD to 0.88SD when comparing sites that have formal decision support to those that did not.
- The developer reports that multiple newer studies have found similar effects. In the first study that compared average SDM scores for breast cancer patients who did and did not use formal decision support, they found statistically significant and higher scores at the practice with decision support (a mean difference of 0.58, p=0.002 at one month and a mean difference 0.61, p=0.0002 at one year). The differences translate to an effect size of 0.43 at one month and 0.51 at one year. The second study compared average SDM scores at an orthopedic practice before and after implementing decision support (a mean difference of 0.2, p=0.009). The difference translates to an effect size of 0.2. The developer notes that the effect size in the orthopedic practice was low due to already using some decision aids and the effect size representing the incremental improvement.
- Overall, the developer suggests that a meaningful difference in scores corresponds to an effect size of at least 0.4 SD.

Missing Data

- In sample four, there were no missing data for the SDM Process score.
- In sample five, one percent of responders skipped one or more items on the scale. Of those with missing responses, two responders skipped all items on the scale and five responders skipped one item on the scale.
- In sample six, the online administration did not allow responders to skip questions, so there were no missing data.
- The developers then compared responses to nonresponders and those with and without missing responses from samples five and six using t-tests or chi square. The developer notes that when

comparing responders and nonresponders, there was no difference between gender, site, or race/ethnicity. However, they did find statistically significant differences by age. Additionally, the developers noted that when comparing those with missing data and without missing data, there were no differences between age, race/ethnicity, gender, clinical topic, or site.

- The developer recommends excluding those with one or more missing responses to the survey, considering missing responses did not have a meaningful impact on scores.

Comparability

- The measure only uses one set of specifications for this measure.

SMP Summary:

- SMP member comments about meaningful differences highlighted the concern that it would be useful to see if an SDM measure can differentiate providers who all practice SDM as opposed to those who do and do not and that the information presented does not fully answer the question.
- There were concerns regarding missing data in that nonresponse bias was not explored fully. Specifically, one SMP member pointed out that in sample 5, mean age for responder was 64.5 and while for non-responder mean age was 59.

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 developer reports that data for this measure is generated online or via paper.
- Electronic patient surveys can be captured via the patient reported outcomes registries or online survey platforms.
- Patient surveys administered via paper are required to be entered into an online database by staff for analysis and reported.
- The developer notes that the only difficulty regarding data collection is in obtaining sufficient high response rates because it requires effort on the part of clinic staff.
- The developer reports there are no fees associated with the use of this measure.

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 notes that the measure is publicly reported to both measured facilities and the public via 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 #2962 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).

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 states that results are provided to measured entities
- The developer states that a free user guide is available which summarizes the psychometrics of the measure, highlight issues regarding implementation and then also clarify scoring.
- The developer notes that the measure development team is not directly involved in data collection, analysis, or feedback.
- The developer reports that feedback was obtained from several stakeholder groups including administrative and clinical leaders in orthopedics, patients, surgeons. They developer states that they meet quarterly with MGB leadership and the PROMs team to track feedback and identify any issues.
- The developer states that the administrative and clinical leaders in orthopedics found the measure acceptable and easy to incorporate into the PROMs set. The developer notes that ability to tie the administration into the existing online patient-reported survey platform was critical to adoption. Based on feedback from orthopedic administrative and clinical leaders, they developers added items to one of the existing time frames to make it easier for the leadership and staff to incorporate. The developer also notes that as a result of the timeframe change only patients who underwent surgery at this time are surveyed, as there is not consistent follow-up with PROMs for non-surgical patients.
- Regarding feedback from patients who have participated in research studies using these measures, the developer states that the measures are highly acceptable to patients with very little missing data. The developer reports no negative feedback with the SDM Process items from respondents and no complaints about undue burden due to these 4-items being added.
- When results were shared with the surgeons, the developer notes that feedback was generally positive and the surgeons felt that survey items are actionable. The developer notes that surgeons often want to see the item-level responses to understand what is or is not driving the scores and/or differences in the scores.
- The developer states that Users have provided feedback on wording of questions which are considered on a case-by-case basis with review of the context in which the items are being used before approving any wording changes.
- The developer states that feedback was used to update the user guide where advice was given to users on how to best set up the survey to ensure high response rates and high quality data. The developer notes that the main advice was to incorporate the survey items into existing registries or patient survey platforms supported by electronic medical records but notes that they are unable to do that at this time.

***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 state that higher scores have also been associated with less decisional conflict and less decision regret reported by patients and that scores improve after the introduction of formal decision support programs.

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 Measures

- NQF #0005: CAHPS Clinician & Group Surveys (CG-CAHPS) Version 3.0 -Adult, Child
- NQF #3227: CollaboRATE Shared Decision Making Score

Harmonization

- NQF #0005: CAHPS Clinician & Group Surveys (CG-CAHPS) Version 3.0 -Adult, Child
    - The developer states that the endorsed CAHPS measures (PCMH and ACO versions) include an optional supplement of shared decision making items, which were adaptations of the items from the SDM Process measure.
    - The developer states that this measure cannot be integrated into the CAHPS protocols due to sample sizes and sample designs.
    - The developer states that this measure approaches sampling by targeting patients who have undergone a procedure rather than the number of ambulatory patients from a clinician's practice or a clinical site used with CAHPS measures which may be small.
    - The developer states that the approach used for NQF #2962 provides the ability to control the sample sizes of respondents and provides for collecting data about the same decision when

using the data to compare clinical sites—which is essential in order to meaningfully interpret the results as measures of quality of care.

- NQF #3227: CollaboRATE Shared Decision Making Score
    - The developer states that NQF #3227 asks patients to rate the quality of the provider's communication but does not capture specific, concrete behaviors that occurred in the clinical encounter like NQF #2962. In addition, the measure does not target a specific clinical decision but rather is intended to be administered following any clinical encounter.

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 included evidence from three samples, a cross-sectional study, a randomized controlled trial and a large prospective sample collected as part of routine care at a large health system.

[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:

Shared decision making (SDM) is a process in which clinicians meaningfully engage patients in medical decisions. SDM involves helping patients recognize that there is a choice to be made, ensuring patients understand the pros and cons of the options and incorporating what matters most to patients into the final choice. As described in Barry et al 2018 NEJM Catalyst, the goal of shared decision making is to improve decision quality, ensuring that decisions are well informed and reflect patient goals, concerns and preferences. SDM has also been associated with lower decisional conflict and less decision regret.

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:

When faced with a medical problem for which there is more than one reasonable approach to treatment or management, shared decision making means providers should outline for patients that there is a choice to be made, discuss the pros and cons of the options and make sure that patients have input into the final decision. The result will be decisions that align better with patient goals, concerns and preferences. This measure asks patients who had any of 7

preference sensitive surgical interventions to report on the interactions they had with their providers when the decision was made to have the surgery.

[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:

The development of the survey was based on conceptual framework of shared decision making first outlined in 1990 and extended in 2000s (See Mulley 1990; Sepucha and Mulley 2003; Sepucha and Mulley 2009). The items were developed with significant input from patients and clinicians (Valentine et al 2021). Patient feedback emphasized the importance of hearing about the options and of talking about both benefits and potential harms. The cognitive interviews also underscored the importance of separating the items that asked about discussion of benefits from the items asking about discussion of harms. Patients also emphasized the importance of having clinicians ask about and listen to patients' goals. Clinicians have remarked that they appreciate that items are specific and focus on measurable behaviors, which means that the results provide actionable feedback for things they can do immediately to improve scores. Administrators and quality officers also appreciate the actionable nature of the items and the lack of ceiling effect which is so common to many patient-reported communication surveys.

Mulley AJ. Methodological issues in the application of effectiveness and outcomes research to clinical practice. In: Heithoff KA, Lohr K, eds. Effectiveness and Outcomes in Health Care. Washington (DC): National Academy Press; 1990.

Sepucha KR, Mulley AG. Extending decision support: preparation and implementation. Patient Educ Couns. 2003; 50(3):269–71.

Sepucha K, Mulley AGJ. A perspective on the patient's role in treatment decisions. Med Care Res Rev. 2009;66(1 suppl):53S–74S.

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. 2021 Feb;41(2):108-119. doi: 10.1177/0272989X20977878. Epub 2020 Dec 15. PMID: 33319648.

[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 meta-analysis of the SDM Process scale for surgical decisions found that SDM Process scores were associated with higher decision quality, less decisional conflict (as measured by higher SURE scores), and lower decision regret (Valentine et al 2021). Further, another published study, focused on hip and knee replacement and spine surgery decisions, found that SDM Process scores were related to less regret and higher patient satisfaction (Brodney et al. 2019).

1. 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. 2021 Feb;41(2):108-119. doi: 10.1177/0272989X20977878. Epub 2020 Dec 15. PMID: 33319648.

2. 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:

When physicians provide balanced information to patients (often in the form of decision aids) and have a discussion about the options and about what patients want, patients answers these questions in a way that reflects a shared decision making process.

[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]

We have collected a great deal of data from surveys of patients who have made decisions drawn from the general population and from clinical sites documenting that for many decisions, patients routinely do not perceive that they discuss the cons of proposed interventions, are not told about alternatives and are not asked to share their treatment preferences as part of the decision. Consistently, their levels of knowledge of information relevant to the decisions they are making are low. We then have evidence that when clinicians commit to shared decision making, by routinely providing decision aids for example, the scores of patients with respect to knowledge and the decision making process are higher. We believe the use of the Shared Decision Making Process Score and appropriate measures of patient knowledge can be catalysts to routinely informing and involving patients in important medical decisions, which in turn will increase the likelihood that patients will get the care they want and that is consistent with their goals and concerns (Stacey et al 2020).

Stacey D, Légaré F, Boland L, Lewis KB, Loiselle MC, Hoefel L, Garvelink M, O'Connor A. 20th Anniversary Ottawa Decision Support Framework: Part 3 Overview of Systematic Reviews and Updated Framework. Med Decis Making. 2020 Apr;40(3):379-398. doi: 10.1177/0272989X20911870. PMID: 32428429.

[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:

The sample includes patients from three different sources (Sample 4, 5, and 6). Sample 4 was a randomized trial (DECIDE-OA study) and includes data from 568 patients with hip or knee osteoarthritis surveyed about 6 months after orthopedic surgeon visit from 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 includes patients (n=646) surveyed by mail within 6 months after hip or knee replacement surgery across a large health system with four main hospitals (two academic medical centers and two community hospitals).

Sample 6 includes data collected across a large health system as part of their orthopedic patient reported outcomes registry (n=5,330) patients surveyed as part of routine care shortly after hip or knee replacement surgery or back surgery across a large health system with four main hospitals (two academic medical centers and two community hospitals) from 2018-2022. The data cover four sites (two academic medical centers and two community hospitals) with 94 surgeons and 5,330 patients. The mean across all decisions was 2.95 and individual site scores ranged from 2.8 (hip/knee) to 3.4 (site 3 herniated disc).

Table 1b.02. SDM Process scores by site for the different clinical topics for Samples 4, 5 and 6

Sample 4	*	*	*	*	*	*	*	*
*	N	Knee M (SD)	N	Hip M (SD)	*	*	*	*
Site 1	59	2.7 (1.1)	50	2.5 (0.8)	*	*	*	*
Site 2	95	2.6 (1.1)	72	2.5 (1.1)	*	*	*	*
Site 3	195	2.55 (1.1)	97	2.4 (1.2)	*	*	*	*
Overall	349	2.6 (1.1)	219	2.4 (1.1)	*	*	*	*
Sample 5	*	*	*	*	*	*	*	*
*	N	Knee M (SD)	N	Hip M (SD)	N	Herniated Disc M (SD)	N	Spinal Stenosis M (SD)
Site 1	66	2.5	66	2.4 (1.1)	22	3.3 (1.1)	55	3.1 (1.1)
Site 2	55	2.5	83	2.6 (1.2)	12	3.5 (1.0)	59	3.0 (1.0)
Site 3	55	2.8	44	2.5 (1.0)	15	3.2 (0.9)	23	3.3 (1.2)
Site 4	17	2.9	14	2.1 (1.2)	26	3.0 (1.2)	27	3.0 (1.1)
Overall	193	2.6 (1.2)	207	2.5 (1.1)	75	3.2 (1.1)	164	3.1 (1.1)
Sample 6	*	*	*	*	*	*	*	*
*	N	Knee M (SD)	N	Hip M (SD)	N	Herniated Disc M (SD)	N	Spinal Stenosis M (SD)
Site 1	649	2.8 (1.0)	562	2.8 (1.0)	173	3.3 (0.9)	292	3.2 (1.0)
Site 2	523	2.8 (1.1)	532	2.8 (1.1)	269	3.3 (0.8)	411	3.0 (1.0)
Site 3	645	2.9 (1.0)	565	2.9 (1.0)	202	3.4 (0.8)	265	3.1 (1.0)
Site 4	130	2.85 (1.1)	112	2.8 (1.1)	n/a	n/a	n/a	n/a
Overall	1947	2.9 (1.0)	1771	2.8 (1.0)	644	3.3 (0.9)	968	3.1 (1.0)

**Cells intentionally left empty*

Previous Submission:

See answer to 1b.3 below

[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:

The new data shows fairly good scores for the orthopedic topics from sites that have been making an effort to engage patients in shared decision making as part of routine care. There is some room for improvement in scores, particularly for hip and knee replacement surgery decisions.

Previous Submission:

We have data comparing clinical sites that have made a commitment to do shared decision making with the shared decision making process scores in usual care, derived both from cross-section surveys of patients who have made the decisions or clinical sites that were making no special effort to implement shared decision making. The data in the five tables in the attachment consistently show that clinical sites that made a special effort to implement shared decision making have "significantly" higher shared decision making process scores from their patients than patients in "usual care".

These data are presented in detail in NQF table attachment

[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:

We have new data to examine disparities for 4 of the 7 decisions using Samples 4-6. Samples 4 & 5 include age, education, race/ethnicity and gender. Sample 6 includes age and gender. Younger respondents and males appear to have slightly higher scores, though most results are neither statistically nor clinically significant. Of note, Sample 6 is large and most characteristics do reach statistical significance; however, the magnitude of the differences in SDM Process scores is quite small (0.1-0.2). The findings do not support disparities for education or race/ethnicity, but the samples are also quite small for race/ethnicity.

Table 1b.04. Samples 4-6 data on disparities by population and clinical topic

Hip replacement surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
VARIABLE	GROUP	N	SDM SCORE M	P	N	SDM SCORE M	P	N	SDM SCORE M	P

Hip replacement surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
AGE	<65	113	2.5	0.34	95	2.5	0.86	643	3.0	0.00
*	65+	106	2.4	*	112	2.5	*	1128	2.7	*
GENDER	FEMALE	129	2.4	0.74	113	2.5	0.90	989	2.7	0.00
*	MALE	90	2.5	*	94	2.5	*	782	2.95	*
RACE	NON-HISPANIC WHITE	203	2.4	0.89	196	2.5	0.97	n/a	*	*
*	OTHER RACES	13	2.5	*	11	2.5	*	n/a	*	*
EDUCATION	COLLEGE GRAD	151	2.5	0.54	n/a	*	*	n/a	*	*
*	NOT COLLEGE GRAD	67	2.4	*	n/a	*	*	n/a	*	*
Knee replacement surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
VARIABLE	GROUP	N	SDM SCORE M	P	N	SDM SCORE M	P	N	SDM SCORE M	P
AGE	<65	156	2.7	0.20	76	2.6	0.73	596	3.0	0.004
*	65+	193	2.5	*	117	2.6	*	1351	2.8	*
GENDER	FEMALE	191	2.55	0.58	107	2.6	0.78	1187	2.8	0.001
*	MALE	158	2.6	*	86	2.6	*	760	3.0	*
RACE	NON-HISPANIC WHITE	319	2.55	0.03	180	2.6	0.42	n/a	*	*
*	OTHER RACES	21	3.1	*	13	2.9	*	n/a	*	*
EDUCATION	COLLEGE GRAD	199	2.6	0.68	n/a	*	*	n/a	*	*
*	NOT COLLEGE GRAD	145	2.6	*	n/a	*	*	n/a	*	*
Herniated Disc Surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
VARIABLE	GROUP	N	SDM SCORE M	P	N	SDM SCORE M	P	N	SDM SCORE M	P
AGE	<65	n/a	*	*	64	3.25	0.33	418	3.4	0.00
*	65+	n/a	*	*	11	2.9		226	3.2	*
GENDER	FEMALE	n/a	*	*	35	3.2	0.83	295	3.2	0.02
*	MALE	n/a	*	*	40	3.2		349	3.4	*

Hip replacement surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
RACE	NON-HISPANIC WHITE	n/a	*	*	71	3.2	0.29	n/a	*	*
*	OTHER RACES	n/a	*	*	4	3.75	*	n/a	*	*
EDUCATION	COLLEGE GRAD	n/a	*	*	n/a	*	*	n/a	*	*
*	NOT COLLEGE GRAD	n/a	*	*	n/a	*	*	n/a	*	*
Spinal stenosis Surgery	*	*	Sample 4	*	*	Sample 5	*	*	Sample 6	*
VARIABLE	GROUP	N	SDM SCORE M	P	N	SDM SCORE M	P	N	SDM SCORE M	P
AGE	<65	n/a	*	*	55	3.2	0.44	281	3.2	0.01
*	65+	n/a	*	*	109	3.0	*	687	3.1	*
GENDER	FEMALE	n/a	*	*	74	3.0	0.73	448	3.0	0.002
*	MALE	n/a	*	*	90	3.1	*	520	3.2	*
RACE	NON-HISPANIC WHITE	n/a	*	*	145	3.1	0.59	n/a	*	*
*	OTHER RACES	n/a	*	*	19	2.9	*	n/a	*	*
EDUCATION	COLLEGE GRAD	n/a	*	*	n/a	*	*	n/a	*	*
*	NOT COLLEGE GRAD	n/a	*	*	n/a	*	*	n/a	*	*

**Cells intentionally left empty*

Previous submission

We have data comparing reported shared decision making process scores by patient age, gender, education and race. Although there are some examples of significant relationships in the data, they do not go in consistent directions. The takeaway from the data is that we do not have evidence that the processes of decision making with providers are consistently related to any of those patient demographic characteristics.

These results are presented in detail in NQF table attachment

[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]

See 1b.4

[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]

N/A

[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]

Shared Decision Making Process

[Response Ends]

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

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

[Response Begins]

This measure assesses the extent to which health care providers actually involve patients in a decision-making process when there is more than one reasonable option. While we believe that the survey will work for patients who have undergone any elective surgical procedure, we have proposed a limited set of surgeries based on existing data for these conditions. This measure focuses on patients who have undergone one of 7 common, important surgical procedures: total hip or knee replacement for osteoarthritis, lower back surgery for lumbar spinal stenosis or herniated disc, radical prostatectomy for prostate cancer, mastectomy for early stage breast cancer or percutaneous coronary intervention (PCI) for stable angina. Patients answer four questions (scored 0 to 4) about their interactions with providers about the

decision to have the procedure, and the measure of the extent to which a provider or provider group is practicing shared decision making for a particular procedure is the average score from their responding patients who had the procedure.

[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]

Cancer: Breast

Cancer: Prostate

Cardiovascular: Coronary Artery Disease (PCI)

Musculoskeletal: Joint Surgery

Musculoskeletal: Low Back Pain

Musculoskeletal: Osteoarthritis

Surgery: Cardiac Surgery

Surgery: Neurosurgery / Spinal

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

Inpatient/Hospital

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]

https://mghdecisionsciences.org/tools-training/sdm-process-survey/

[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]

Patient answers to four questions about whether or not 4 essential elements of shared decision making (laying out options, discussing the reasons to have the intervention, discussing reasons not to have the intervention, and asking for patient input) are scored and summed. A group/practice score is the average of their patient scores.

[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]

All responding patients will answer four questions about their pre-surgical interactions with their providers:

1. How much did a doctor (or health care provider) talk with you about the reasons you might want to (HAVE INTERVENTION)—a lot, some, a little, or not at all?
2. How much did a doctor (or other health care provider) talk with you about reasons you might not want to (HAVE INTERVENTION)—a lot, some, a little or not at all?
3. Did any of your doctors ask you if you wanted to (HAVE INTERVENTION)? (YES/NO)
4. Did any of your doctors (or health care providers) explain that you could choose whether or not to (HAVE INTERVENTION)? (YES/NO) OR: "Did any of your doctors (or health care providers) explain that there were choices in what you could do to treat your [condition]? (YES/NO)

SCORING: 1 POINT EACH FOR ANSWERING "A LOT" OR "SOME" TO QUESTIONS 1 AND 2; 1 POINT EACH FOR ANSWERING "YES" TO QUESTIONS 3 AND 4. TOTAL SCORE = 0 TO 4.

The score for a provider group is simply the average score for their responding patients. This will be a continuous number from 0 to 4.

[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]

While we believe that the survey will work for patients who have undergone any elective surgical procedure, we have proposed a limited set of surgeries based on existing data for these conditions.

All responding patients who have undergone one of the following 7 surgical procedures: back surgery for a herniated disc; back surgery for spinal stenosis; knee replacement for osteoarthritis of the knee; hip replacement for osteoarthritis of the hip; radical prostatectomy for prostate cancer; percutaneous coronary intervention (PCI) for stable angina, and mastectomy for early stage breast cancer.

[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]

See S2. There is an attached Excel file with ICD 10 and CPT codes needed to identify eligible patients. A published manuscript describes the development and validation of an algorithm using ICD 10 and CPT codes that can be used to identify eligible orthopedic 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]

For back, hip, knee, and prostate surgery patients, there are no exclusions as long as the surgery is for the designated condition (for example, hip replacement for osteoarthritis not for hip fracture).

For PCI, we are focused on patients who are treated for stable coronary artery disease. As such, those who had a heart attack within 4 weeks of the PCI procedure are excluded, as are those who have had previous coronary artery procedures (either PCI or CABG).

For mastectomy, we are focused on females having mastectomy as the primary surgical treatment for breast cancer. Patients who had had a prior lumpectomy for breast cancer in the same breast, patients who have not been diagnosed with breast cancer (who are having prophylactic mastectomies), and males with breast cancer are excluded.

Respondents who are missing one or more responses to the SDM Process measure do not receive a total score and thus, are excluded.

[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]

Included in attached Excel file. A published manuscript describes the development and validation of an algorithm using ICD 10 and CPT codes that can be used to identify eligible orthopedic 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.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]

none

[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]

Continuous variable, e.g. average

[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]

Better quality = Higher score

[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]

All responding patients will answer four questions about their pre-surgical interactions with their providers:

1. How much did a doctor (or health care provider) talk with you about the reasons you might want to (HAVE INTERVENTION)—a lot, some, a little, or not at all?
2. How much did a doctor (or other health care provider) talk with you about reasons you might not want to (HAVE INTERVENTION)—a lot, some, a little or not at all?
3. Did any of your doctors ask you if you wanted to (HAVE INTERVENTION)? (YES/NO)
4. Did any of your doctors (or health care providers) explain that you could choose whether or not to (HAVE INTERVENTION)? (YES/NO) OR: "Did any of your doctors (or health care providers) explain that there were choices in what you could do to treat your [condition]? (YES/NO)

SCORING: 1 POINT EACH FOR ANSWERING "A LOT" OR "SOME" TO QUESTIONS 1 AND 2; 1 POINT EACH FOR ANSWERING "YES" TO QUESTIONS 3 AND 4. TOTAL SCORE = 0 TO 4.

The score for a provider or provider group is simply the average score for their responding patients. This will be a continuous number from 0 to 4.

[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 one of the 7 target procedures for the target indication are identified from medical records, claims or in some other way. Patients can be sampled sequentially, or a pool of such patients who had the procedure in a particular time period can be created and sampled at a rate that produces the desired number of potential respondents.

These same questions 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 measures have been used in surveys using both of those models. However, the basic proposal here is to use the measure to evaluate clinical care provided by particular clinical sites, provider groups, or providers.

With respect to sample sizes, the standard deviations vary some by procedure. For most procedures, comparing samples of size of 50 or larger will detect differences of .5 in Decision Process Scores ($p < .05$), which is an order of magnitude we have often observed between sites that do and do not make an effort to do shared decision making. Samples of 100 reduce that number to around .3. We think samples in the range of 50 to 100 offer sufficient power to detect clinically meaningful differences in clinical practice.

[Response Ends]

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

[Response Begins]

Proxy respondents are not permitted. Virtually all of the patients who receive these procedures should be able to answer survey questions. We think it is important to get the perceptions of the patients themselves about the process.

[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]

We have administered the survey by mail, telephone and online. Reminders, either by mail, phone or email, and small incentives are helpful to increase the rate of response. Data collection protocols must make it clear that individual answers will not be viewed by the physician and/or his/her staff. Results from similar surveys have made it clear that survey responses are skewed if respondents think they can be reviewed by their providers or clinic support staff. Therefore, we recommend that data not be collected from respondents who are in a clinic or hospital setting.

Calculate response rate as all those responding divided by all those invited to answer the survey questions (AAPOR response rate 4). We recommend that data not be accepted if response rates are lower than 50%.

[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]

We have used these questions in mail, telephone and online surveys. 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 4: DECIDE-OA data (n=568) patients with hip or knee osteoarthritis surveyed about 6 months after orthopedic surgeon visit from 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: Cross-sectional survey of patients (n=646) by mail within 6 months after hip or knee replacement surgery across a large health system with four main hospitals (two academic medical centers and two community hospitals).

Sample 6: Data collected across a large health system as part of their orthopedic patient reported outcomes registry (n=5,330) patients surveyed as part of routine care shortly after hip or knee replacement surgery or back surgery across a large health system with four main hospitals (two academic medical centers and two community hospitals).

Previous Submission:

1) TRENDS, a national survey of adults over 40 in Knowledge Networks panel who had made decisions were used to estimate "usual care" experience for back, knee and hip decision experience (2012)

2) Surveys of Medicare patients who had mastectomy, prostate cancer surgery or PCI were used to estimate usual care experiences for those procedures (2008)

3) Demonstration site data. Nearly 3000 patients were surveyed in 6 different clinical sites around the US that were implementing the use of decision aids and encouraging shared decision making for 14 different decisions about testing and surgery. These data were collected from 2009 through 2013. The numbers varied by procedure and are presented in the appropriate tables. These data were used to assess the decision making process scores for patients in setting in which clinical sites were making an effort to implement shared decision making.

They were also used to estimate the reliability of average Shared Decision Making Process scores for clinical sites. Most of the usable data for that analysis came from Dartmouth medical center, because they had the most responses, and we wanted 20 or more responses in each random half estimate of the rating at a practice for a particular decision. The analysis was based on responses from 663 patients over 5 different decisions.

In addition, we had data from 4 clinical sites from 266 patients who made decision about breast cancer treatment, 1 site emphasizing use of DAs and shared decision making and the other three in "usual care" mode which we used to add to our validity data.

[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 4) 04-19-2016 - 2-28-2018

Sample 5) 07-05-2018 – 12-07-2018

Sample 6) 08/06/2018 – 05/21/2022

Previous Submission:

2008 to 2014

[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 4: Participants were from three orthopedic practices each at three different sites: one academic medical center, one community hospital and one specialty orthopedic hospital in the Northeast.

Sample 5: Participants were from four orthopedic practices within one large health system, two practices each at an academic medical center, and two practices each at a community hospital in the Northeast.

Sample 6: Participants were from four orthopedic practices within one large health system, two practices each at an academic medical center, and two practices each at a community hospital in the Northeast.

Previous Submission:

Although we used our cross-sectional data from surveys for estimates of usual care values, including means and SDs, all of the evidence for the values related to the validity and reliability of the measure to reflect clinical practice represents average patient reported scores, either for individual practices or a combination of practices that either were or were not making a special effort to promote shared decision making

[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 4: The sample included n=568 follow-up surveys. 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 646 patients (73% response rate) who had recently undergone hip or knee replacement surgery or spine surgery with 36 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 65 years old (SD 11years), 51% female, 93% White, non Hispanic, and 63% had hip or knee replacement surgery (versus spine surgery). The sample is described in more detail in Valentine et al 2021.

Sample 6: The sample includes 5,330 patients from 88 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 years old (SD 11), 55% were female and 70% had knee or hip replacement surgery (versus spine surgery).

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 Submission:

The cooperating practices all varied in the decisions for which they used decision aids and how they were distributed. The data on the Shared Decision Making Process in demonstration sites were mainly collected by sending out a mail questionnaire after patients had met with providers about the decisions. Response rates varied by site and decision. We usually included all patients who completed a questionnaire.

[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 6 is used for reliability. Samples 4, 5 and 6 were used for validity.

Previous Submission:

(Not answered)

[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 Social risk factors for samples 4, 5, & 6

Characteristic	Sample 4 (N=568)	Sample 5 (N=639)	Sample 6 (N=5,330)
Gender: Female, n (%)	320 (56%)	329 (51%)	2931 (55%)
Age, mean (SD)	65 (9)	65 (11)	67 (11)
Race/Ethnicity: White, non Hispanic, n (%)	522 (94%)	592 (93%)	n/a
Education: ≥ College graduate, n (%)	350 (62%)	n/a	n/a
Joint: n (%)	*	*	*
Hip	219 (39%)	209 (33%)	1771 (33%)
Knee	349 (61%)	209 (33%)	1947 (37%)
Spinal Stenosis	n/a	137 (21%)	968 (18%)
Herniated Disc	n/a	84 (13%)	644 (12%)

**Cells intentionally left empty*

Previous Submission:

Profile of all patients respondent in the demonstrations sites:

Respondent characteristics	n (%)
Age group (n = 2,928)	*
<50	438 (15)
50 – 64	1,533 (52)
≥ 65	957 (33)
Gender (n = 2,961)	*
Male	1,881 (63)
Female	1,080 (37)
Education (n = 2,914)	*
High school or less	926 (32)
Some college or 2-y college	819 (28)
4-year college or more	1,169 (40)
Race (n = 2,832)	*
White	2,721 (96)
Black	68 (2)
Ethnicity: (n = 2,893)	*
Hispanic	62 (2)
Non Hispanic	2831 (98)

**Cells intentionally left empty*

[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 randomly split patients making the same decision at the same clinical site into groups of 58 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. In all, we had 76 patient groups created from 5,294 patient reports.
2. We calculated the intra-class correlation coefficient (proportion of the total variance accounted for by the between-site variance).

Previous Submission:

1. At the item level, we measured test-retest reliability from same individuals 4 weeks apart
2. At the item and score levels for an encounter, we compared patient reports with coding of tape recordings of encounters
3. At the practice level, we randomly split patients making the same decision 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.

[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. For reliability at the level of clinical practice, we divided patients from the same site making the same decision into random groups and correlated their Process Scores. With minimum sample sizes of 58, we created 76 patient groups from 5294 patient reports. Site 1 (back condition) = 10 groups, Site 1 (condition hip/knee) = 14 groups; Site 2 (back condition) = 8 groups, Site 2 (condition) hip/knee) = 16 groups; Site 3 (back condition) = 8 groups, Site 3 (condition hip/knee) = 16 groups; Site 4 (condition hip/knee) = 4 groups. We get an average reliability of 0.69 95% CI (0.685, 0.69).
2. Intraclass correlation by dividing between site variance by total variance, ICC=0.96.

Previous Submission:

The Decision Process Score is technically a composite, with conceptual roots in what a good decision process should look like, so a calculation of Cronbach's alpha may not be an appropriate measure of reliability (see Bollen and Lennox, 1991), but we have calculated them for some decisions, and they are reasonably high (often in the .5 to .7 range)

We have short term (~4 weeks) test-retest data on some variations of this measure and obtained ICC values in the .7 to .8 range.

We also have two tests of whether or not patient reports of their interactions align with coding of tape recordings of the same interactions. In one study, objective observers and patients both rated various aspects of the interactions between doctors and patients making breast cancer decisions. The results showed a high level of agreement, although patients' ratings tended to a bit higher, on average, than observers'(Pass et al, 2012).

In a different test, women's interactions with physicians about primary treatment for breast cancer were tape recorded (n = 96). Coding of the interactions were related to patient reports using the questions in the Process Score. In this case, because the clinically reasonable options were known, questions were asked separately for discussion of the pros and cons of both reasonable options. Kappas were computed for the dichotomous variables and product moment correlations for the multi-category items between the coded results and what respondents said. For the overall scores, the correlations were .50 ($p<.001$) and .38 ($p=.004$) for adjuvant therapy and surgery decisions respectively. With respect to individual items, the values were higher for whether the options were presented (.64 to .71) and how much the reasons for each option were discussed (.64 to .75) and lower for how much the cons were discussed (.16 to .46) and whether the patient's input was sought (.14 to .32).

Finally, for reliability at the level of clinical practice, we have divided patients from the same site making the same decision into random groups and correlated their Process Scores. With minimum sample sizes of 25, we get an average reliability of .61. The numbers would be higher with larger samples, which we hope to have soon.

1. Bollen K, Lennox R. Conventional wisdom on measurement: A structural equation perspective. *Psychol Bull*. 1991;110(2):305-314. http://psycnet.apa.org/psycinfo/1992-03966-001.
2. Pass M, Belkora J, Moore D, Volz S, Sepucha K. Patient and observer ratings of physician shared decision making behaviors in breast cancer consultations. *Patient Educ Couns*. 2012;88(1):93-99. doi:10.1016/j.pec.2012.01.008.

[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:

As expected, with the larger sample size, the reliability at the practice level did improve for the scores. The results indicate reasonably consistent average scores.

Previous Submission:

We think the reliability of the overall process is satisfactory at both the individual encounter level and at the clinical practice level. In particular, at the practice level, which is the level that is more relevant for the way we propose to use this measure, we think the reliability will only get higher with bigger samples.

[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)

Accountable Entity Level (e.g. hospitals, clinicians)

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:

Clinician/Site-Level Analysis:

Two published studies provide data at the site level. Fowler et al 2021 summarizes performance of the measure at the group/practice level. We tested whether or not clinical practices that were making a special effort to implement shared decision-making had higher SDM Process scores than sites practicing "usual care" or than cross-sections of patients who faced the same decision. We used t-tests to compare mean SDM Process scores from different settings, using a Welch's correction when needed. We also calculated Cohen's d effect sizes for all comparisons. This effect size indicates the difference between groups in terms of their standard deviations where a d of 0.2 would indicate a small effect, a d of 0.5 indicates a medium effect and a d of 0.8 indicates a large effect.

Content validity:

The Valentine 2021a paper presents data on the development process including cognitive interviews, conceptual framework, and content validity.

Patient-Level Analyses:

Three published studies provide additional evidence supporting validity of the measure (Brodney et al 2019, Valentine et al. 2021a and Valentine et al 2021b). Construct validity was examined through hypothesis testing regarding predicted relationships with other decision making and health outcomes:

1. We hypothesized that patients who have higher SDM Process scores would also have higher confidence (as measured by the SURE scale, a short form of the decisional conflict scale), higher satisfaction, and less regret.
2. We tested hypotheses that higher SDM Process scores are associated with higher rates of Informed, Patient-Centered surgery.
3. We tested the hypothesis that higher SDM Process scores were associated with better health outcomes post-surgery.

For the Valentine et al 2021a paper, secondary meta-analysis was conducted across 8 studies of 11 surgical conditions to identify the validity of the scale. First, overall effect size was calculated by employing the inverse variance method and random effects model and the heterogeneity of effects was calculated using the DerSimonian–Lair estimator of between-study variance.

For the Brodney 2019 paper, we used generalized linear and logistic regression models with the General Estimating Equations approach to account for clustering of patients within surgeons. Models adjusted for patient characteristics such as age, gender, education, joint, and baseline quality of life scores.

For the Valentine et al. 2021b paper, we used generalized linear and logistic regression models with the General Estimating Equations approach to account for clustering of patients within surgeons in a cross-sectional sample to identify relationships between the scale and health outcomes

References:

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

Fowler FJ Jr, Sepucha KR, Stringfellow V, Valentine KD. Validation of the SDM Process Scale to Evaluate Shared Decision-Making at Clinical Sites. J Patient Exp. 2021 Nov 26;8:23743735211060811. doi: 10.1177/23743735211060811. PMID: 34869847; PMCID: PMC8640277.

Valentine KD, Cha T, Giardina JC, Marques F, Atlas SJ, Bedair H, Chen AF, Doorly T, Kang J, Leavitt L, Licurse A. Assessing the quality of shared decision making for elective orthopedic surgery across a large healthcare system: cross-sectional survey study. BMC musculoskeletal disorders. 2021a Dec;22(1):1-0.

Valentine KD, Vo H, Fowler Jr FJ, Brodney S, Barry MJ, Sepucha KR. Development and evaluation of the shared decision making process scale: a short patient-reported measure. Medical Decision Making. 2021b Feb;41(2):108-19.

Previous Submission:

We have a number of ways we have looked at validity. The approach is described below with each individual approach to testing.

[Response Ends]

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

Examples may include correlations or t-test results.

[Response Begins]

Current Submission:

Clinician/Site-Level Analysis:

For osteoarthritis of the knee and hip, patients in the practices where decision aids were used reported significantly better decision processes than a cross-section sample of adults who faced the same decisions (2.9 vs. 2.5, P < .001, d=0.49 and 2.9 vs. 2.1, P < .001, d= 0.84 respectively). The difference in SDM Process Scores for spine practices that did and did not use decision support (3.0 vs. 2.75, P=.12, d=0.22) was in the expected direction but was not large enough to reach statistical significance. In breast cancer, the practice that had formal decision support had significantly better scores than cancer practices without any decision support interventions (2.7 vs 2.3, P< .05, d=0.47).

Content validity: Extensive cognitive testing of the items was done with patients (N = 88) across 17 different clinical conditions between 2006 and 2009. Some key insights included the limited ability of patients to adequately describe a shared decision (i.e., even when patients labeled a decision shared, their narrative descriptions were not consistent with a shared process) and their general desire to rate their clinicians highly. We found that asking patients to rate or evaluate provider behavior or their interactions proved problematic because patients lacked a frame of reference for evaluating the quality of a decision process. These findings resulted in the SDM Process survey's focus on a specific clinical decision and on reports of events or behaviors, rather than ratings or evaluations of their clinicians or of a particular clinical interaction. We believe this has been critical in reducing the problems with ceiling effects that occur for other patient reported engagement surveys.

Patient-Level:

1. From Valentine et al 2021, a meta analysis across 11 surgical contexts (n=3,965) found SDM Process scores were related to higher decision confidence (effect size d=0.57, p<0.001), and lower decision regret (effect size d= -0.34, p <0.001). From Brodney et al 2019 (n=649), we have evidence that SDM Process scores were higher among patients who reported no regret (2.5 (1.2) no regret vs. 2.3 (1.2) regret , p<0.001 for hip and knee surgery) and for patients who reported high satisfaction (2.3 (1.2) not satisfied vs. 2.5 satisfied (1.2), p<0.001) for hip and knee surgery) and (2.1 (1.4) not satisfied vs. 2.6 (1.2) satisfied , p<0.001) for back surgery.
2. From Valentine et al 2021, a meta-analysis across 11 surgical contexts (n=3,965) found SDM Process scores were related to higher rates of Informed, Patient-Centered decisions (effect size d=0.18, p=0.03). From Brodney et al (2019), SDMP scores were also significantly higher for patients who made informed, patient centered decisions compared to those who did not, for hip and knee replacement surgery (2.7 vs. 2.3, p<0.001) and for back surgery (3.2 (0.9) vs. 2.0 (1.3), p<0.001).
3. From the Valentine, Vo, et al, 2021 analyses we found higher SDM Process scores were associated with larger improvements from pre- to post-surgery in mental health (b=0.16, p=0.02) and physical health (b=0.25, p=0.02) outcomes for patients who had total joint replacement of the hip or knee for osteoarthritis, but not patients who had spine surgery (all ps>0.26).

Previous Submission:

The evidence for the value of clinical practices devoted to shared decision making and that the SDM Process score is a valid measure of clinical performance comes from a number of studies of decision making in clinical practices, some of which were trying to implement shared decision making on a routine basis and using decision aids for many decisions. The following summarizes those results.

We have compared the aggregate SDM Process Scores from patients treated clinical sites that have committed to shared decision making, usually by including the routine use of decision aids, with reports of national cross-sections of patients from the TRENDS survey who made the same decisions.

Table 2. Mean SDM Decision Process Scores at SDP Demonstration sites and from a national sample of patients for three orthopedic procedures.

Data Source	Decision Topic	N	Mean Process Score	Std. Deviation
TRENDS	Surgery: Knee Pain	163	2.81	1.139
Demo Sites	Knee Osteoarthritis	239	3.24**	.840
TRENDS	Surgery: Hip Pain	57	2.45	1.236
Demo Sites	Hip Osteoarthritis	129	3.31***	.864
TRENDS	Surgery: Low Back Pain	152	3.23	1.016
Demo Sites	Herniated Disc + Spinal Stenosis	55	3.38	.828

**p < .01

*** P< .001

For osteoarthritis of the knee and hip, it can be seen that the patients in practices where decision aids are used reported significantly better decision processes than a cross-section sample of adults who faced the same decisions. The responses did not differ for conversations about lower back pain, but the decisions about back pain were by far the best decision processes based on respondent reports in the national survey.

Because the data in the above table were collected with quite different time periods between the decision and the measurement, a better test may come from studies of breast cancer decision making in four clinical sites. One of these four sites routinely used decision aids and had support for patients when they met with their surgeons to facilitate getting patients' questions asked and answered. The other three sites practiced usual care, with no special intervention to encourage shared decision making.

Table 3. Mean Decision Process Scores from a SDP demonstration site, three "usual care" sites and a cross-section sample of Medicare patients for decision for how to treat breast cancer

Data source	N	Mean Process Score (SD)	t (comparing with demonstration site)	p-value
SDP Demonstration site	40	3.00 (.934)	*	*
Usual care sites	227	2.54 (1.205)	2.7	<.01
Survey of Medicare beneficiaries treated for breast cancer	914	1.85 (1.25)	3.7	<.001

*cell intentionally left blank

Table 3 shows that the SDP demonstration site patients reported a decision process that was much better than those clinical sites where there was no intervention to promote decision making. The comparable data from the survey of Medicare patients describing their decision making process for breast cancer treatment were much lower still.

Table 4 shows similar data for decision making around hip and knee replacement.

Table 4. Mean Decision Process Scores from a SDP demonstration and three "usual care" sites and a cross-section sample of adults who made decisions for how to treat arthritis of the hip or knee.

Data source	N	Mean Process Score (SD)	t (comparing with demonstration site)	p-value
SDP Demonstration site	178	2.96 (1.04)	*	*
Usual care sites	204	2.6 (1.06)	3.3	<.001
TRENDS National survey of adults who made decisions about knee or hip replacement	268	2.70 (1.17)	2.5	<.02

*cell intentionally left blank

As in Table 3, we see in Table 4 that the SDP demonstration sites had significantly better process scores from their patients than sites with no shared decision making initiative and was better than the national sample reported as well.

Finally, a small study at a clinical site in Stillwater, Minnesota collected data using the SDM Process Score questions from patients who discussed treatment for benign prostatic hyperplasia (BPH) with their urologists. They started collecting these data before introducing decision aids and continued to collect them after the use of decision aids that encouraged shared decision making became routine in the practice. Table 5 shows the results. While the SDM Process Score was pretty good before the use of decision aids, it was significantly better after they were introduced.

Table 5. Mean Decision Process Scores before and after the introduction of decision aids into process of treatment decisions for BPH.

When data collected	N	Mean Process Score (SD)	t (comparing before and after data)	p-value
Before use of decision aids	47	3.02 (.794)	3.12	<.01
After use of decision aids began	16	3.63 (.619)	*	*

*cell intentionally left blank

[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 development process with extensive cognitive interviewing highlighted the importance of anchoring patients on a specific decisions and on reporting specific behaviors rather than general and abstract ratings of their physicians or an interaction. The resulting items were understood as intended and acceptable for patients to answer. At the group level, the SDM Process scores can detect differences between sites that do and do not have formal decision support services. At the patient-level the SDM Process scores demonstrated construct validity and are associated with less decisional conflict, higher satisfaction and less regret. Further, in certain symptom-driven conditions, higher SDM Process scores were associated with better mental and physical health outcomes. The data suggest that spine surgery is strong in terms of the amount of SDM present in discussions. Other elective surgical decisions demonstrate considerable room for improvement.

Previous Submission:

In summary, we have data that show clearly that decision making on average in the US as measured by the SDM Process score is not very good and that clinical sites that commit to improved decision making attain average scores from their patients that are much higher than average. We think this is one of relatively few instances in which outcome measures based on patient reports are clearly linked to the way that clinical practices are trying to interact with patients.

[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 examined differences between site-level scores with multivariable linear regression analyses with Generalized Estimating Equations to correct for correlated error due to patients being nested within surgeons.

Previous Submission:

As noted in the analyses above, we simply used t tests to assess differences between mean Shared Decision Process scores from patient who made decisions in practices that had implemented procedures to promote shared decision making and patients who made decisions in usual care, either based on data from practices or from our national surveys.

[Response Ends]

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

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

[Response Begins]

Current Submission:

As data reported in the previous submission, studies using the SDM Process survey have found effect sizes ranging from 0.39SD – 0.88SD when comparing sites that have formal decision support (coaching or decision aids) and those that did not. Since then, new data has found similar effects. For example, Sepucha et al 2019 used generalized linear models with generalized estimating equations to account for clustering of patients within breast surgeons to compare average SDM Process scores for breast cancer practices that did and did not use formal decision support and found statistically significant and higher scores at the practice with decision support (mean difference 0.58 [SE 0.18], p=0.002 at 1 month) and that persisted at 1 year (mean difference =0.61 [SE 0.18], p=0.002. (Sepucha et al 2017). This difference translates to an effect size of 0.43 and 0.51 at 1 month and 1 year respectively. Another study (Sepucha et al. 2017) compared scores at the same orthopedic practice before and after implementing decision support, using repeated measures analysis with generalized estimating equations and found a mean difference of 0.2 (SD 1.1) p=0.009 in SDMProcess scores. This result translates into an effect size of 0.2. This effect size is low, in part because the orthopedic practice was already engaging in some use of decision aids (27% of patients were receiving decision aids before the project and that increased to 64% after) and this effect size reflects that incremental improvement. Based on these analyses, we suggest that a meaningful difference in scores corresponds to an effect size of at least 0.4SD.

Sepucha KR, Langford AT, Belkora JK, Chang Y, Moy B, Partridge AH, Lee CN. Impact of Timing on Measurement of Decision Quality and Shared Decision Making: Longitudinal Cohort Study of Breast Cancer Patients. Med Decis Making. 2019 Aug;39(6):642-650. doi: 10.1177/0272989X19862545. Epub 2019 Jul 29. PMID: 31354095; PMCID: PMC7240785.

Sepucha K, Atlas SJ, Chang Y, Dorrwachter J, Freiberg A, Mangla M, Rubash HE, Simmons LH, Cha T. Patient Decision Aids Improve Decision Quality and Patient Experience and Reduce Surgical Rates in Routine Orthopaedic Care: A Prospective Cohort Study. J Bone Joint Surg Am. 2017 Aug 2;99(15):1253-1260. doi: 10.2106/JBJS.16.01045. PMID: 28763411.

Previous Submission:

The practices using decision aids and promoting shared decision making consistently had significantly better scores on this measure. The exception is decisions about surgery for lower back pain, which consistently get very high scores in "usual care". We think that is not a reflection of a problem with the measure but a reflection of the way back surgery decisions are made.

[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:

The evidence supports that many practices have fairly low scores if they do not have formal decision support services and this measure is able to pick up meaningful differences between groups/practices that do and do not engage patients in shared decision making for surgical decisions. Further, it is able to detect statistically significant differences over time for practices that may be implementing interventions to improve or enhance their ability to engage patients in shared decision making.

Previous Submission:

We think the evidence is pretty strong that this measure validly reflects the extent to which a clinical practice is practicing shared decision making.

[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:

Recent use of the surveys continues to find very little missing data. In Sample 4, there were no missing data for the SDM Process score (0/568) and in Sample 5, 1% of responders (7/646) skipped one or more items on the scale. Of those with missing responses, 2 responders skipped all items on the scale and 5 responders skipped one item on the scale. In sample 6, the online administration did not allow respondents to skip items so there were no missing responses. We compared responders to non responders (for sample 5 and 6) and those with and without missing responses using t-tests or chi square. As described in the previous submission, excluding those with one or more missing responses to the survey is reasonable approach to handle missing data.

Previous Submission:

In our experience, it is relatively rare for respondents to the surveys not to answer all four questions. For example, from our test sites from patients making decisions about knee replacement and lower back surgery for herniated disc and spinal stenosis, the percentages of respondents not answering all four questions were 3%, 5% and 0% respectively; the percentages having more than one missing response were <1%, 2% and 0% respectively, from a total of 411 respondents. We did not experiment with alternative ways of handling missing data, because it really could not affect the results. We think leaving out anyone not answering all the questions or imputing a .5 score for one missing response (and eliminating anyone with more than one missing answer) would both be reasonable approaches to dealing with missing data.

[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:

When comparing responders to non responders, we did not find differences by gender, site or race/ethnicity. We did find statistically significant differences by age. Responders tended to be older, but the difference was not large (mean age 67 (SD 12) vs. 66 (SD 14), $p<0.001$). When comparing those with missing data to those without missing data, we did not find any differences by age, race/ethnicity, gender, clinical topic or site.

We considered two approaches to handling missing data, excluding anyone with a missing item or imputing a .5 score for one missing response and excluding anyone with more than one missing answer. Table 2b.09a shows the frequency of missing data and Table 2b.09b shows the results of these two approaches for the one sample with missing data.

Table 2b.09a Frequency of missing data

Number of SDMP questions answered	Frequency (%)	SDMP score excluding missing	SDMP score imputing 0.5 for missing
0	2 (0.3%)	n/a	n/a
1	0 (0%)	n/a	n/a
2	0 (0%)	n/a	n/a
3	5 (1%)	n/a	2.7
4	639 (99%)	2.75	2.75

Table 2b.09b. Responder and Non responder analyses for Samples 5 and 6.

*	Sample 5 Responder	Sample 5 Non Responder	Sample 5 p-value	Sample 6 Responder	Sample 6 Non Responder	Sample 6 p-value
N	647	242	*	7203	9958	*
Female N (%)	335 (48%)	123 (49%)	0.86	3911 (54%)	5331 (54%)	0.33
Male N (%)	312 (52%)	119 (51%)	*	3292 (46%)	4627 (46 %)	*
Age Mean (SD)	64.5 (11)	59 (14)	0.00	67 (12)	66 (14)	0.00
White, non Hispanic	605 (93%)	219 (90%)	0.16	n/a	n/a	*
Non White	42 (6.5 %)	23 (9.5 %)	*	n/a	n/a	*
Site 1	212 (33%)	68 (28%)	0.32	68 (28%)	212 (33%)	0.32
Site 2	212 (33%)	85 (35%)	*	85 (35%)	212 (33%)	*
Site 2	85 (13%)	41 (17%)	*	41 (17%)	85 (13%)	*
Site 3	138 (21%)	48 (20 %)	*	48 (20%)	138 (21%)	*

*Cell intentionally left blank

Previous Submission:

See 2b.08 above.

[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:

There were minimal differences between responders and non responders. Missing data for responders was very low and did not differ by patient characteristics (age, gender, or race/ethnicity). Due to the small amount of missing data, the

approach to handling missing responses did not have meaningful impact on scores. As a result, we recommend that any survey missing one or more responses be excluded from the analyses.

Previous Submission:

See 2b.08 above

[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 mentioned in the previous submission, in defining the sample of patients who are clinically appropriate to receive the survey, there are inclusion and exclusion criteria. For example, that the hip replacement is for treatment of osteoarthritis, the mastectomy is primary treatment for female breast cancer and that the PCI is for treatment of stable coronary artery disease. We did not send surveys to patients with the exclusion codes, as a result, we do not have any data to test relating to those codes.

We also recommend excluding responses for those who miss one or more of the SDM Process items. To evaluate the effect of exclusions due to missing data across groups, we examined the frequency of included and excluded responses for each site and tested for difference using a chi-square test.

Previous Submission:

The exclusions for these measures only apply to two of the decisions. For mastectomy, we exclude males and those who have had previous surgery for breast cancer because that may indicate a situation where there are not reasonable medical alternatives. We exclude prophylactic mastectomy because it is not treating cancer. For PCI, we exclude those who had a recent heart attack, because there is enough evidence of life extension that the decision may be seen as skewed toward the intervention. Those who have had previous coronary artery interventions may also be in complex medical situations. What we are trying to do is restrict measure to those with stable angina, which is a condition for which there clearly are alternatives to PCI and for which the paradigm of shared decision making clearly applies. Thus, the exclusions are specifically targeted to focus on those patients for whom shared decision making is clearly appropriate. We did not test effects of clinical exclusions on measures because they are about the clinical appropriateness of the measure.

[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:

Seven respondents with one or more items missing were excluded. When we examine the impact of these exclusions, we find they have a negligible impact on performance measure scores mainly due to the small number of exclusions.

Table 2b.17 Comparing those with and without excluded data

*	No Missing	Missing SDM Process score	p-value
N	639	7	*
Age <65 N (%)	290 (45%)	2 (29%)	0.47
White, non-Hispanic N (%)	592 (93%)	6 (86%)	0.42
Female N (%)	329 (51%)	5 (71%)	0.45
Site 1 N(%)	209 (33%)	3 (43%)	0.59
Site 2 N(%)	209 (33%)	3 (43%)	*
Site 3 N(%)	84 (13%)	1 (14%)	*
Site 4 N(%)	137 (21%)	0 (0%)	*
Hip N (%)	207 (32%)	2 (29%)	0.83
Knee N (%)	193 (30%)	3 (43%)	*
Sinal Stenosis N(%)	164 (26%)	1 (14%)	*
Herniated Disc N(%)	75 (12%)	1 (14%)	*

*Cell intentionally left blank

Previous Submission:

Not applicable, no formal testing of exclusions.

[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:

Overall, we had very few exclusions. We did not find significant or meaningful differences by site or patient characteristics, though the power to detect differences was limited due to the very small number of exclusions.

[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]

There are two reasons we are not recommending any kind of risk adjustment. First, and perhaps most important, there is no ethical basis for saying the standards for engaging in shared decision making for these preference-sensitive surgical decisions should vary by patient characteristics. Second, as the data on disparities in section 1.4. Tables 6-10 in the NQF_table_attachment shows, we have not found any meaningful or systematic differences in average shared decision making scores based on age, gender, education or ethnicity. So, in our experience, adjustments would not have any meaningful effect on results.

[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 staff at sites will need to enter the patient data into an online database for analysis and reporting.

[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. This approach could be easily replicated in other online patient survey platforms.

[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]

Current submission: 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. 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 people interested in administering the survey items used to generate the measure, provided the survey is 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: BlueCross Blue Shield of Massachusetts Alternative Quality Contract

URL: see either https://www.bluecrossma.org/aboutus/our-mission or https://coverage.bluecrossma.com/article/shes-guardian-angel

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 Shared Decision Making Process survey
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 (see below for more details)
3. By the end of 2022, they are planning to collect data directly from their members using the SDMProcess measure as a basis for confidential reporting to providers, with the eventual 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 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 program requires practices to provide a description of the process for assessing the quality of shared decision making using the decision quality assessment tool (that includes the Shared Decision Making Process items). 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. The QPP is looking for reporting capability and evidence of process implementation. 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: MassGeneralBrigham Shared Decision Making Program

URL: https://www.massgeneralbrigham.org/en/about/newsroom/articles/shared-decision-making

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 SDM Process measure into the Patient Reported Outcomes Registry. All patients undergoing primary hip or knee replacement surgery and spine 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 SDM Process measure 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 the use of measures that don't add value. Contracts between payers (commercial and Medicaid) and provider organizations 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 SDM Process measure is currently in use in the Blue Cross Blue Shield of MA Alternative Quality Contract. It is also been approved as part of the aligned measure set available to entities in MA to be used in commercial and Medicaid contracts.

[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 in use.

[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/sdm-process-survey/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 regarding the overall scores and scores by procedure, by surgeon, and by site quarterly. The team present to the department leadership once 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. The program is still collecting data. Our team is not directly involved in data collection, analysis or feedback.

[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 measure acceptable and easy to incorporate into the PROMs set. Patients have not complained about undue burden due to these 4-items being added. The ability to tie the administration into the existing online patient-reported survey platform was critical to adoption. Further, we added the items to one of the existing time frames (as opposed to creating an additional survey) and this also made it easier for the leadership and staff to incorporate. 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 not had any negative feedback with the SDM Process items from respondents. When we have shared results with the surgeons, we have had generally positive feedback. They appreciate that the items are actionable and often want to see the item-level responses to understand what is or is not driving the scores and/or differences in the scores. When comparing across procedures, the gaps become clear. For example, we have found that spine surgeons are much more likely to discuss non surgical options and reasons not to have surgery compared to hip and knee replacement surgeons. (see Valentine et al 2021)

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.

[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. Many users want the survey to be able to be used in other clinical areas.

[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. We would be very interested in expanding the set of clinical areas and have considerable data on its use in medication and cancer screening decisions, however, we were unable to pull together all of the required components to expand the specifications for this submission.

[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]

As described earlier, quality improvement studies have shown that scores improve after the introduction of formal decision support programs (such as patient decision aids or decision coaching services). Further, higher scores have also been associated with less decisional conflict and less decision regret reported by patients. Engaging patients in medical decisions is a key component of patient-centered care and critical to ensuring that patients are well informed and receive preferred treatments. Documenting SDM Process scores provides actionable information for groups and practices to help their clinical teams achieve high quality, patient-centered care.

[Response Ends]

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

[Response Begins]

Since the implementation of the measure, we have 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 data and the Informed Patient Centered Hip/Knee Replacement (#2958) measure as part of that work. Another surgeon is interested in looking at the SDM scores throughout the COVID pandemic to examine the quality of decisions in virtual visits compared to in-person visits. A third surgical colleague is going to look at whether use of decision aids and SDM Process scores are associated with better outcomes for patients from disadvantaged communities.

[Response Ends]

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

[Response Begins]

Other than the examples mentioned above, we have also been approached by other colleagues who are interested in adapting the measure for new contexts (for example, we have been working with genetic counseling and the National Society of Genetic Counselors) to adapt and evaluate a SDM Process survey regarding genetic testing decisions. Further, BlueCross BlueShield MA is also interested in generating sufficient data to support extending specifications for the measure to cover cancer screening and common medication decisions.

[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]

0005: CAHPS Clinician & Group Surveys (CG-CAHPS) Version 3.0 -Adult, Child

3227: CollaboRATE Shared Decision Making Score

[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]

American Academy of Hip and Knee Surgeons Performance Measure Set (https://www.aahks.org/practice-resources/performance-measures/)

Measure #3: Percentage of patients undergoing a total hip replacement with documented shared decision-making including discussion of conservative (non-surgical) therapy (e.g. NSAIDs, analgesics, exercise, injections) prior to the procedure

Measure #2: Percentage of patients undergoing a total knee replacement with documented shared decision-making including discussion of conservative (non-surgical) therapy (e.g. NSAIDs, analgesics, exercise, injections) prior to the procedure

[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]

The endorsed CAHPS measures (PCMH and ACO versions) include an optional supplement of shared decision making items, which were adaptations of the items from the SDM Process measure. The CAHPS measures were used for respondents who reported they had discussed starting or stopping a prescription medication (for PCMH) and for patients who reported discussion a prescription medication or a procedure with a provider (ACO). The problems with integrating this measure into the CAHPS protocols relate to both sample sizes and sample designs. This measure works best when applied to a specific kind of decision (eg. Decision to take medication for high blood pressure or decision to have surgery for herniated disc.) CAHPS samples relatively small numbers of ambulatory patients from a clinician's practice or a clinical site. Those samples do not include enough encounters at which decisions are made about specific medications or specific tests or surgical procedures to provide reliable data. Hence, they had to ask about any decisions about starting or stopping medications or surgical procedures and combine the answers for each type of decision. The numbers of such decisions tend to be very small, even when all medications or procedures are combined. Moreover, we have abundant data showing that the Shared Decision Making Process Score varies widely from medication to medication and procedure to procedure. (Zikmund=Fisher et al, 2010; Fowler et al, 2012; Fowler et al, 2014). The approach we are proposing, sampling patients who have undergone a procedure, provides the ability to control the sample sizes of respondents and provides for collecting data about the same decision when using the data to compare clinical sites—which is essential in order to meaningfully interpret the results as measures of quality of care.

3227 CollaboRATE: The endorsed CollaboRATE measure of shared decision making assesses the patient's perception of how much effort was made in 1) making sure patients understood their health issue, 2) listening to issues that matter to the patient, and 3) including those issues in choosing next steps. However, this measure asks patients to rate the quality of the provider's communication and does not capture specific, concrete behaviors that occurred in the clinical encounter. In addition, the measure does not target a specific clinical decision but rather is intended to be administered following any clinical encounter. Without targeting, the general samples do not include enough encounters at which decisions are made about specific medications or specific tests or surgical procedures to provide reliable data. Further, scores may combine ratings for different types of decisions such as a test, medication or surgical procedure . This is also potentially problematic as we have abundant data showing that the Shared Decision Making Process Score varies widely from decision type (screening to medication to procedure and then vary widely depending on the type of procedure). We believe it is essential to identify the decision in order to understand and act on the results.

The AAHKS measures rely on surgeon documentation of shared decision making in the visit note. The data collection burden is less than patient-reported measures (though it will require chart review for each case). Further, the measure lends itself to simply using a templated note in order to meet the criteria, as opposed to having a meaningful conversation that engages and informs patients.

References:

1. Zikmund-Fisher BJ, Couper MP, Singer E, Ubel PA, Ziniel S, Fowler FJ Jr, Levin CA, Fagerlin A. Deficits and variations in patients' experience with making 9 common medical decisions: the DECISIONS survey. Med Decis Making. 2010 Sep-Oct;30(5 Suppl):85S-95S. doi: 10.1177/0272989X10380466. PMID: 20881157.
2. Fowler FJ Jr, Gallagher PM, Bynum JP, Barry MJ, Lucas FL, Skinner JS. Decision-making process reported by Medicare patients who had coronary artery stenting or surgery for prostate cancer. J Gen Intern Med. 2012 Aug;27(8):911-6. doi: 10.1007/s11606-012-2009-5. Epub 2012 Feb 28. PMID: 22370767; PMCID: PMC3403150.
3. Fowler FJ Jr, Gerstein BS, Barry MJ. How patient centered are medical decisions?: Results of a national survey. JAMA Intern Med. 2013 Jul 8;173(13):1215-21. doi: 10.1001/jamainternmed.2013.6172. PMID: 23712194.

[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]

While not directly competing as the target populations are not overlapping, SDM Process and Collaborate are quite similar in intent. The SDM Process scores focus on concrete, observable behaviors in the visit, and across a wide range of clinical areas, have not shown evidence of floor or ceiling effects that are common with the other endorsed measures. The three individual Collaborate items are each rated on a 10-point scale, yet due to very high ceiling effects, the CollaboRATE scale is reported as a top score (the percent top score vs anything less), limiting its power to detect differences. A study directly comparing the psychometric performance of three SDM measures (including SDM Process and CollaboRATE) found that both measures had evidence of predictive validity in the hypothesized direction for decision regret and satisfaction; however the Shared Decision Making measure performed better than CollaboRATE in discriminating among patients who did and did not review a decision aid and it also was associated with higher decision quality (where Collaborate was not). Further, this study found that the measures were correlated (Pearson correlation 0.38 for hip and knee and 0.41 for backs, $p<0.001$) but not too strongly, suggesting they are measuring different aspects of care. (Brodney et al. 2019)

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.

[Response Ends]