NATIONAL QUALITY FORUM
Driving measurable health improvements together

Memo

June 22, 2020

To: NQF Patient Experience and Function Standing Committee

From: NQF staff

Re: Corrections to Preliminary Analysis form for NQF 3559

Dear Patient Experience and Function Standing Committee,

The Yale CORE team has pointed out that some sections of their submission were cut off from the preliminary analysis form of NQF 3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA). This memo will provide supplementary inclusions that you may find helpful in your assessment of the measure. Also included are the developer's responses to Scientific Methods Panel concerns and a summary of a submitted analysis related to disparities.

Omissions of Submitted Materials on Preliminary Analysis Form

Page 43: For "S.2b. Data Dictionary, Code Table, or Value Sets" the attachment provided is the Testing Attachment, but for this section should be the data dictionary: HipKneePROPMDataDict_ForSubmission.xlsx

Page 43: S.2d. repeats the question from S.2c. and not the correct question from the Submission Form, which is "If this is an instrument-based measure, please indicate responder."

Page 61: Under 2b1.2, the "Empirical Measure Score Validity", some text was dropped off the submission. Dropped text in *italics*:

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

[OMITTED TEXT FOLLOWS]: *We examined the distribution of THA/TKA PRO-PM RSIRs by THA/TKA RSCR national categories within hospitals submitting complete PRO data for at least 25 THA/TKA procedures. Like the THA/TKA PRO-PM we are seeking to validate, NQF #1550 measures outcomes of elective primary THA/TKA procedures. While the outcomes of these two measures are not clinically expected to be perfectly correlated, they both reflect hospital-level quality of care for patients experiencing elective primary THA/TKA surgery. It is clinically accepted that certain THA/TKA complications, specifically mechanical complications that require further surgical interventions, revision or even removal of the*

prosthetic joint, are associated with worse patient-reported outcomes such as pain and function. Therefore, we would anticipate that hospitals with higher overall risk-adjusted complication rates in this population (patients undergoing elective primary THA/TKA procedures) would see overall worse patient-reported outcomes. Higher complication rates (RSCRs) would lead to worse clinical outcomes such as increased pain and decreased function, resulting in lower percentages of patients achieving a substantial clinical benefit (lower RISRs). Thus, in terms of measure results, we expect an inverse association between the RSIRs of the THA/TKA PRO-PM and the RSCRs of Risk-Standardized Complication measure elective primary THA/TKA.

Our examination of this association between RSIRs and RSCRs used categories of hospital performance on the THA/TKA complications measure:

- *Hospitals worse than national average (those with higher complication rates),*
- *Hospitals the same as national average, and*
- *Hospitals better than national average (those with lower complication rates).*

With this approach, we expect that hospitals within the RSCR category "Worse than the National Average" will have lower risk-standardized improvement rates (RSIRs) and that hospitals within the RSCR category "Better than National Average" will have higher RSIRs.

Page 63 – 64: Under 2b1.4, the "Empirical Measure Score Validity". Dropped text in *italics*.

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

[OMITTED TEXT FOLLOWS]: *As perfect or even high correlations are not expected given the different time periods and cohorts, we sought to show the conceptual relationship between the two outcomes through actual hospital-level measure results (RSIRs) grouped by statistically significantly categories of performance of RSCRs.*

Page 89 – 90: "For 4.1. Current and Planned Use" the incorrect planned use is identified. We reported that the planned use would be "a. Public Reporting" but the Measure Worksheet has Payment Program listed. (Please see HipKneePROPMSubNQFForm_For_Submission_Updated_1-30-2020.docx.)

Clarification on Disparities

Yale CORE has also suggested that the Committee would benefit from the summarization of a disparity-related analysis within the staff summary of disparities:

Page 4, Disparities: Correlation coefficients between RSIRs calculated without social risk factors in the risk model and RSIRs calculated individually for each of the social risk factors (dual eligibility, non-White race, and AHRQ SES Index indicate near perfect or perfect correlation in our data.

Responses to Scientific Methods Panel Concerns

The staff preliminary analysis refers to a document the developer put together in response to concerns identified by the NQF Scientific Methods Panel. It is provided in its entirety below:

Developer Response to Scientific Methods Panel's Preliminary Analysis

Measure Number: 3559

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

Measure Developer/Steward: Yale New Haven Health Services Corporation - Center for Outcomes Research and Evaluation (YNHHSC - CORE)/Centers for Medicare & Medicaid Services (CMS)

Reliability

- **Issue 1:** Measure specifications: Some NQF Panel members wanted clarification on the measure result calculation and definitions for "predicted," "expected" and "overall observed" improvement.
    - **Developer Response 1:** A description of the approach to measure score calculation, including a definition of each of these terms, is provided in Section 2b3.1.1 of the NQF Testing Attachment. We estimated the hospital-specific Risk-Standardized Improvement Rate (RSIR) using a hierarchical logistic regression model (hierarchical model). We calculate the hospital-specific RSIRs as the ratio of a hospital's "predicted" number of improvements to "expected" number of improvements multiplied by the overall observed improvement rate. This approach is analogous to a ratio of "observed" to "expected" that people may be familiar with. It conceptually allows for a comparison of a hospital's performance given its case-mix to an average hospital's performance with the same case-mix.

      Hospital-level RSIR Calculation =

      $$\frac{\text{Predicted Improvement}}{\text{Expected Improvement}} \times \text{Observed Overall Improvement Rate}$$

    - The expected number of cases meeting SCB improvement for each hospital (denominator) was estimated using the hospital's patient mix and the *average* hospital-specific intercept (the average intercept among all hospitals in the sample). The expected SCB improvement for each patient was calculated via the hierarchical model (HLM formula provided in NQF Testing Attachment, Section 2b3.1.1), which applies the estimated regression coefficients to the observed patient characteristics and adds the *average* hospital-specific intercept. Operationally, the expected number of cases meeting SCB improvement for each hospital was obtained by summing the expected improvement of all elective primary THA/TKA patients in the hospital.
    - The predicted number of cases meeting SCB improvement for each hospital (numerator) was estimated using its patient mix and an *estimated* hospital-specific intercept. The predicted improvement for each patient was calculated via the hierarchical model, which applies the estimated regression coefficients to the observed patient characteristics and adds the hospital-specific intercept. The predicted number of cases meeting SCB improvement for each hospital was calculated by summing the predicted improvement of all elective primary THA/TKA patients in the hospital.

- The overall observed improvement rate is the unadjusted overall rate of SCB improvement for all patients across all hospitals.

- **Issue 2:** There was a request for clarification about how the measure accounts for patients that die between the hospital discharge and the postoperative PRO data collection period (270-365 days postoperatively), and if they are considered "lost to follow-up." Another NQF Panel member noted that excluding deaths seemed reasonable but suggested a check on death as a possible adverse event.

    - **Developer Response 2:** Patients who do not provide postoperative PROM scores (at 270 to 365 days following surgery) are not counted in the denominator or the numerator of the measure because they have incomplete PROM data. Presently, this includes patients who expire between the time of hospital discharge and the postoperative assessment window. The measure denominator is primary elective THA/TKA patients and therefore there is a lower than average competing mortality rate for this group of patients. Deaths within 30 days of the procedure are already captured in CMS' THA/TKA complications measure, with which this measure is fully harmonized. However, we will continue to work with CMS to monitor mortality and its impact on measure validity.

- **Issue 3:** An NQF Panel member noted that the HOOS, JR and KOOS, JR appeared to have been transformed from 0-100 but no specifications on the approach to transformation were provided.

    - **Developer Response 3:** Scoring of HOOS, HR and KOOS, JR are exactly as specified by the instrument developer. Stephen Lyman and colleagues[1],[2] scaled the HOOS, JR and KOOS, JR to 100 points (as was done with the original HOOS and KOOS instruments), with 0 representing total hip or total knee disability and 100 representing perfect hip or knee health, respectively. Scores for the HOOS, JR and KOOS, JR were determined using Rasch-based person scores from each instruments' validation cohort. A crosswalk table provided by the authors for the HOOS, JR and KOOS, JR converts raw sum scores to the interval level measure scaled from 0 to 100. The HOOS, JR and KOOS, JR scores were derived from the responses to full HOOS surveys from both registries.

- **Issue 4:** An NQF Panel member noted that the interval over which the "change" in score appears to have been estimated (90-0 days prior to surgery and 270-365 days following surgery) is quite wide and could vary for an individual patient by as much as 6 months.

    - **Developer Response 4:** The Technical Expert Panel (TEP) considered both data assessment timeframes very carefully. When considering the preoperative assessment window, the TEP believed that elective primary THA and TKA candidates were unlikely to have significant changes in preoperative PROM scores within 90 days of surgery. In addition, they indicated that the additional time to collect data would increase response rates and likely better represent stable and complete recovery from either procedure. Likewise, the postoperative assessment window

---

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

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

was considered very carefully. After considerable input from TEP members, public comments, a thorough literature review, and a review of registry experiences, we defined the postoperative PROM data collection timeframe to between 270 days and 365 days. The TEP concurred with this recommendation. This timeframe allows for full recovery from both THA and TKA and increases opportunity for PRO response.

- **Issue 5:** An NQF Panel member noted concern about attributing changes in joint function to the hospital (versus care such as rehabilitation services) with a follow-up interval of nine months to one year following surgery.
    - **Developer Response 5:** The goal of this hospital-level PRO-PM is to capture the full spectrum of care to incentivize collaboration and shared responsibility for improving patients' health and reducing the burden of their disease. As per our response on the NQF Evidence Form, Section 1a.2.**,** we note that optimal clinical outcomes for patients undergoing an elective primary THA or TKA depend not just on the surgeon performing the procedure, but also on the entirety of the team's efforts in the care of the patient, care coordination across provider groups and specialties; and the patients' engagement in their recovery. Even the best surgeon will not get outstanding results if there are gaps in the quality of care provided by others caring for the patient before, during, and/or after surgery. Further, the hospital has significant influence over discharge and rehabilitation planning for its surgical patients.
- **Issue 6:** An NQF Panel member noted that this appears to be a composite measure, but that NQF form does not appear to have been completed.
    - **Developer Response 6:** This is not a composite measure. The outcome measure is not a composite of a THA PRO-PM and a TKA PRO-PM. Instead, the cohort for this measure consists patients undergoing an elective primary total hip or total knee arthroplasty, and outcomes for patients in the cohort are determined using a single risk model.
- **Issue 7:** Some NQF Panel members had questions about the 25 case volume threshold—what the threshold was based on, what happens to a facility that falls below the 25-case recommendation, if facilities without 25 cases would be excluded from the measure (and should be identified as an exclusion), and if excluded, whether it would create an incentive for them to not complete data.
    - **Developer Response 7:** A 25 case volume threshold is consistent with volume thresholds used for public reporting of claims-based measures with which this measure was intentionally harmonized. It is <u>not</u> a measure exclusion; rather, the recommendation is that hospitals that perform fewer than 25 elective primary THA or TKA procedures during the measurement period or have complete PRO data on fewer than 25 THA or TKA procedures during the measurement period not be included in public reporting of the measure. This recommendation is made to address concerns about the reliability of measure results for hospitals with a small number of procedures and/or procedures with PRO data. And the aggregate number of elective primary THA/TKA procedures conducted among hospitals performing fewer than 25 of these procedures is small; while 33% of hospitals conducted fewer than 25 elective primary THA and TKA procedures from July 1, 2016 to June 30, 2017, the procedures performed

at these hospitals represented just 3.14% (11,175 of 333,850) of the total number of elective primary THA and TKA procedures performed across all hospitals.

- It is expected that hospitals with fewer than 25 procedures total or fewer than 25 procedures with complete PRO data would still receive hospital specific reports, informing them of measure results, but that a Risk-Standardized Improvement Rate (RSIR) for these hospitals would not be publicly reported. Hospital-specific reports provided to these hospitals might also positively impact the collection of PRO data even if an RSIR was not publicly reported.

- **Issue 8:** An NQF Panel member had a clarification question about the data used for reliability testing.
  - **Developer Response 8:** The Combined Dataset consists of the hospitals in both the Development and Validation Datasets combined that have at least 25 elective primary THA/TKA Patients with PRO Data. These 123 hospitals make up the Combined Dataset used for reliability and validity testing. This is consistent with the measure as specified with a 25-case volume threshold.

- **Issue 9:** Concern was expressed about data element reliability testing for "critical data elements" other than the HOOS, JR and the KOOS, JR. Data elements of concern were noted to be those that "make-up the denominator," the two additional PRO tools used in the risk model, and, additional risk factors, including the clinical characteristics based on coding (e.g. liver disease, severe infection).
  - **Developer Response 9:** The codes used to define the measure cohort (denominator) are harmonized with CMS' publicly reported, NQF-endorsed hospital-level THA/TKA complications measure. This measure has been in public reporting since 2013 and undergoes annual updates through independent clinical review by orthopedic coding experts to ensure the measure methodology reflects current clinical and coding practice. In addition, we only use data elements in claims that have both face validity and reliability. We do not use fields that are inconsistently coded across providers. We only use fields that are consequential for payment and which are audited. We identify these variables through empiric analyses and our understanding of CMS auditing and billing policies and do not use variables which do not meet this standard. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud, and audits important data fields used in our measures, including diagnosis and procedure codes, and other elements that are consequential to payment. While we have not performed medical record chart review validation of this measure risk model, multiple CMS claims-based measure risk models have been validated using chart review, a few of them cited

here.[3,4,5]

- The risk variables included in this risk model were defined over several years through multiple, iterative steps that pulled in stakeholder input on feasibility, clinical capture, accuracy, reproducibility and clinical face validity. These steps included surveying orthopedic practices regarding the feasibility, uniformity and reliability of risk variables identified by clinical experts and published literature; a consensus summit by orthopedic specialty societies to narrow and prioritize clinical risk variables for prospective collection as part of the CJR model – these recommendations were adopted in toto by CMS; additional clinical and empiric evaluation in CJR data; and by TEP approval.
- Patients and the TEP were engaged throughout the measure development process. The TEP was thoroughly engaged in the selection of risk variables for inclusion in the risk model, providing input on the importance and feasibility of each variable. Both the TEP and the Patient Working Group provided detailed input on the measure outcome definition.

- **Issue 10:** An NQF Panel member noted that developers reported that HOOS JR was not tested for reliability because the HOOS was tested several times, and do not state it was tested here.
  - **Developer Response 10:** As required by NQF, reliability and validity of the HOOS, JR and KOOS, JR are provided in detail in the NQF Testing Attachment. Section 2a2.3 clarifies that internal consistency reliability using the Person Separation Index (PSI) was assessed for the HOOS, JR, as was a principal component analysis, conducted on the standardized residuals and indicating that the six HOOS, JR items existed in a single dimension. Only test-retest reliability for the HOOS, JR was dependent on results from assessment of the HOOS domains from which the HOOS, JR pain and functioning questions were drawn.
- **Issue 11:** An NQF Panel member stated that "Reliability testing appears to have been done at the patient not hospital level (the unit of comparison of the measure)" and later noted that an ICC comparing hospital level results appears not to have been performed.
  - **Developer Response 11:** Per NQF Guidance (for example, see page 2, Note 10 of the NQF Testing Attachment Form as well as check-off box, Section 2a2.1 suggesting the use of signal-to-noise analysis for measure score validity), we conducted hospital-level reliability testing with signal-to-noise analysis for comparison of hospital-level measure scores. Results are provided in Section 2a2.3 of the NQF Testing Attachment.

---

[3] Krumholz H, Normand S, Keenan P, et al. Hospital 30-Day Pneumonia Readmission Measure Methodology [Internet]. Yale New Haven Health Services Corporation/ Center for Outcomes Research and Evaluation;2008. Available at: http://www.qualitynet.org/dcs/ContentServer?c¼Page&pagename¼QnetPublic%2FPage%2FQnetTier3&cid¼1219069855841.

[4] Krumholz H, Normand S, Bratzler D, et al. Risk-Adjustment Methodology for Hospital Monitoring/Surveillance and Public Reporting Supplement#1: 30-Day Mortality Model for Pneumonia [Internet]. Yale University;2006. Available at: http://www.qualitynet.org/dcs/ContentServer?c¼Page&pagename¼QnetPublic%2FPage%2FQnetTier3&cid¼1163010421830.

[5] Keenan PS, Normand SLT, Lin Z, et al. An Administrative Claims Measure Suitable for Profiling Hospital Performance on the Basis of 30-Day All-Cause Readmission Rates Among Patients With Heart Failure. *Cardiovascular Quality and Outcomes*. 2008;1(1):29-37.

- **Issue 12:** An NQF Panel member voiced concern about proxy assessment, noting that it "is unorthodox and can add significant noise."
    - **Developer Response 12:** As this is a measure of elective procedures, proxy assessments are uncommon (in our data, of the 81% of data submissions with respondent identified, only 8.8% were identified as surrogate responses) but we chose to include these patients in order to ensure they were being measured. We will advise CMS to continue to examine these patients in reevaluation.
- **Issue 13:** Two NQF Panel members voiced concern about missing data, and that the only complete data were analyzed without accounting for what is likely "fairly extensive missingness." One of these members noted concern that missing surveys were accounted for but that missing responses within the survey were not.
    - **Developer Response 13:** We provide a detailed accounting in the NQF Testing Attachment in Sections 2b6.1 through 2b6.3 of our approach to PRO non-response (including elective primary THA and TKA patients with no PRO data and patients missing either preoperative or postoperative data or missing or out-of-range values on PRO data submitted). Due to the voluntary nature of PRO survey data and because PRO data are unlikely to be missing at random, we understand that accounting for potential non-response bias is important for this measure. Since bias may be introduced by systematic differences between responders such as patients with different social risk, we included social risk factors and race in the propensity score models used to create stabilized inverse probability weights to address potential non-response bias.
    - On Section S16 of the NQF Submission Form, we note the importance of high response rates for measuring hospital quality with PROs: "High response rates allow PRO-PMs to best represent hospital quality performance. Hospitals and physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Strong leadership support within the hospital, flexibility in rearranging clinical workflows to accommodate PRO data collection, accessibility of PRO data in real-time to inform clinical decision making can all increase staff investment in the value of PROs in improving care and quality, and PRO data used for clinical decisions can increase patient investment."
    - Regarding missing responses within the survey data, in Section 1.7 of the NQF Testing Attachment, we state that only PROs with complete data are used in measure development and testing. [Complete PRO data is defined as the submission of preoperative patient-reported outcome measure (PROM) and risk variable data with no missing or out-of-range values for required data elements and that could be matched to postoperative PROM data with no missing or out-of-range values, for an elective primary THA/TKA procedure identified in claims data for the measurement period.] For the voluntary data collection, missing data are better addressed through accounting for non-response bias than through data imputation. The TEP supported this approach.

Validity

- **Issue 1:** An NQF Panel member suggested that the exclusion of staged procedures might eliminate up to 43% of procedures, and that the measure name should include "from unstaged procedures."

- **Developer Response 1:** Please note that, globally, the prevalence of staged THA/TKA procedures that are not simultaneous and occur within 1 year of each other is roughly 7%.[6,7] We are happy to clarify this in the measure name if the committee feel this is required.
- Among hospitals submitting PRO data, 7.06% of THA and TKA procedures were staged procedures during the measurement period (two or more procedures occurring during the measurement period in distinct hospitalizations).
- As we note in Section 2b2.2 of the NQF Testing Attachment, 491 (4.17%) of patients with complete PRO and risk variable data had staged procedures during the measurement period. Across hospitals, the mean proportion of procedures excluded from the analysis was 3.84% (SD 5.69), and the median proportion was 2.11%.

- **Issue 2:** An NQF Panel member noted concern that data were not provided on how the excluded patients impact the performance measure scores.
  - **Developer Response 2:** Because the assessment of the measure outcome is unclear in patients with staged procedures – that is, it is hard to clarify the impact of the index procedure on the PRO result – we did not include staged procedures in the measure score. Our clinical consultants and Technical Expert Panel agreed with this exclusion. We are happy to recommend to CMS that staged procedures be reexamined during reevaluation. Because this exclusion is based on the inability to appropriately attribute the outcome to the index procedure, we are uncertain how to interpret the results requested by the Panel member.
- **Issue 3:** Concern was expressed about the 25-case volume recommendation: that it is not identified as an exclusion, that this represents 52% of hospitals in the denominator, and that not considering or testing hospitals that fall below this threshold is a major potential threat to the measure's validity unless the denominator is redefined as suggested above.
  - **Developer Response 3:** As noted in Issue #3 for Reliability above, a 25-case volume threshold is consistent with volume thresholds used for public reporting of claims-based measures with which this measure was intentionally harmonized. It is <u>not</u> a measure exclusion; the recommendation is that hospitals that perform fewer than 25 elective primary THA or TKA procedures during the measurement period or have complete PRO data on fewer than 25 THA or TKA procedures during the measurement period not be included in public reporting of the measure. However, these hospitals will receive confidential measure results.
- **Issue 4:** Another NQF Panel member: The model was developed including cases from hospitals not used for reliability, validity and missing data testing, i.e., hospitals with low caseloads (n<25) not recommended for this measure. Did the developers do a sensitivity test to assess the impact

---

[6] Stefánsdóttir A, Lidgren L, Robertsson O. Higher early mortality with simultaneous rather than staged bilateral TKAs: results from the Swedish Knee Arthroplasty Register. *Clin Orthop Relat Res*. 2008;466(12):3066–3070. doi:10.1007/s11999-008-0404-3

[7] Garland A, Rolfson O, Garellick G, Kärrholm J, Hailer NP. Early postoperative mortality after simultaneous or staged bilateral primary total hip arthroplasty: an observational register study from the Swedish Hip Arthroplasty Register [published correction appears in BMC Musculoskelet Disord. 2015;16:263]. *BMC Musculoskelet Disord*. 2015;16:77. Published 2015 Apr 8. doi:10.1186/s12891-015-0535-0

of excluding these hospitals from the risk-adjustment development sample on the risk-adjustment model?

- **Developer Response 4:** The risk model was developed using all cases in the Development Dataset and validated using all cases in the Validation Dataset. The recommendation for a 25-case volume threshold is for public reporting, and therefore reliability and validity analyses were conducted on hospitals with at least 25 elective primary THA and TKA procedures with PRO data. Including all the THA or TKA procedures for the risk model development will maximize the available information and is a commonly accepted approach.

- Analysis to address non-response included all THA and TKA procedures conducted at all 238 hospitals. The hospitals with 25 or more procedures are reported with weighting for non-response, as per the recommendation for a 25-case volume threshold for public reporting.

- **Issue 5:** An NQF Panel requested clarity for the data provided in T.11 and whether there are meaningful differences between hospitals in the top quartile.

  - **Developer Response 5:** Table 11 indicates a range of Risk-Standardized Improvement Rates (RSIRs) from the 75th to the 100th percentile of hospitals of 72.51% to 86.84%. These RSIRs indicate the risk-standardized proportion of patients achieving substantial clinical benefit improvement following elective primary THA or TKA. The 14.3 percentage points representing this range represent a meaningful difference in the proportion of patients experiencing substantial clinical benefit improvement.

- **Issue 6:** An NQF Panel member asked if the impact of IPW on hospital ratings was assessed by conducting a sensitivity analyses?

  - **Developer Response 6:** In the NQF Testing Attachment in Section 2b6.2, Table 14 we provide a comparison of the mean and distribution of hospital RSIRs with and without stabilized inverse probability weighting. As we note in interpretation of results in Section 2b6.3, this comparison reveals only a small impact on the measure results of adjusting for potential non-response. However, we expect that non-response bias will be a factor for the THA/TKA PRO-PM measure, due to associations with non-response including socioeconomic status and health status. We therefore retained response bias adjustment for the measure results.

- **Issue 7:** Two NQF Panel member**s** expressed additional concern about the extent of missing data and the subsequent threat to measure validity (also noted under Reliability, Issue 13 above). One Panel member noted their belief that the proposed solution (stabilized inverse probability weighting) assumes data missing at random and suggested that requiring near-complete data rather than rely on proxies and statistical modeling was the only good solution.

  - **Developer Response 7:** As noted in our response to Issue 13 under Reliability (above), due to the voluntary nature of PRO survey data and because PRO data are unlikely to be missing completely at random, we understand that accounting for potential non-response bias is important for this measure. Since bias may be introduced by systematic differences between responders such as patients with different social risk, we included social risk factors and race in the propensity score models used to create stabilized inverse probability weights to address potential non-response bias.

- Stabilized inverse probability weighting, calculated using propensity model, does not assume that data are missing completely at random; rather, that particular patient groups have different response rates that are accounted for in the weighted model.
- On Section S16 of the NQF Submission Form, we note the importance of high response rates for measuring hospital quality with PROs: "High response rates allow PRO-PMs to best represent hospital quality performance. Hospitals and physicians incorporating PRO data collection into clinical workflows are likely to reap considerably higher response rates. Strong leadership support within the hospital, flexibility in rearranging clinical workflows to accommodate PRO data collection, accessibility of PRO data in real-time to inform clinical decision making can all increase staff investment in the value of PROs in improving care and quality, and PRO data used for clinical decisions can increase patient investment."

- **Issue 8:** An NQF Panel member asked why the overall observed improvement rate would be used both in the development of the HLM as the dependent variable, and then again in the calculation of the RSIR?
  - **Developer Response 8:** The overall observed improvement rate is used in the calculation of the hospital-level RSIR (as noted in Issue #1 under the "Reliability" heading above, it is a constant to assist the interpretation of RSIR and has no material impact on RSIR) but is not the dependent variable of the HLM model. The dependent variable for this model is a patient-level outcome, identifying the individual patient's improvement.
- **Issue 9:** An NQF Panel member asked how health literacy how will be measured in practice.=
  - **Developer Response 9:** In Section 2b3.1.1 of the NQF Testing Attachment, we list the variables included in the final risk model and note that Health Literacy is assessed by response to the Single Item Literacy Screener questionnaire, which asks about "Comfort Filling Out Medical Forms by Yourself"). The response options are noted in the Data Dictionary in Row 5 of the "Risk Variables with PRO Data" tab: 0 = Not at all, 1 = A little bit, 2 = Somewhat, 3 = Quite a bit, 4 = Extremely.
- **Issue 10:** Concern was expressed about data element validity testing, that published validity data from the HOOS, JR and KOOS, JR were provided, but testing for other critical data elements was not provided.
  - **Developer Response 10:** Please see the detailed response to Issue 9 under Reliability above.
- **Issue 11:** A few NQF Panel members noted concern about ceiling effects of the HOOS and KOOS. One member noted that recent publications have supported the use of other non-condition specific measures (e.g. PROMIS physical function) as valid alternatives for future consideration.
  - **Developer Response 11:** Thank you for this input. We will be sure that CMS and the measure reevaluation contractor are provided this suggestion.
- **Issue 12:** There was a question about the logic in selecting a single threshold for SCB (by THA/TKA). It was noted that there is "a wealth of published literature on the dependency of clinically important improvement thresholds on initial scores." It was suggested that patients with worse initial scores would need to see greater improvement to reach "clinically important

improvement thresholds" than patients with higher initial scores. Concern was raised about the Measure Developer's statement that the SCB outcome allows patients with poor baseline PRO scores to improve, that some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB, and that this biases the measure and may not meet a patient's expectations of improvement. Also, concern was expressed that this approach would penalize providers with higher performing patients at admission.

- **Developer Response 12:** With strong TEP support, this PRO-PM measures improvement with a threshold for the HOOS, JR and for the KOOS, JR tested by Stephen Lyman and colleagues[8] (developers of the HOOS, JR and KOOS, JR) and identified using an anchor-based question to assess substantial clinical benefit (SCB, 22 points for HOOS, JR and 20 points for KOOS, JR). This improvement threshold approach to the outcome was preferred over alternatives (averaging change among patients, measuring a postoperative average or minimum state, or a combined approach of improvement and postoperative state). An improvement threshold approach was preferred for the following reasons:
  - It measures improvement only and discourages surgeons from performing THA/TKA procedures on patients with milder symptoms, as patients with high preoperative PROM scores cannot statistically meet the improvement threshold;
  - It equally rewards hospitals performing THA/TKA on patients with moderate and severe symptoms, as it does not define an "end state" that patients must achieve, only substantive improvement from where they started;
  - Avoids creating what is known as a ceiling effect, where many patients can meet the outcome criteria and decreases the ability of the measure to identify performance variation; and
  - It has less risk of unintended consequences. Specifically, we were concerned that requiring patients to meet a postoperative minimum symptom state would encourage hospitals and their surgeons to avoid offering THA/TKA surgery to anyone with severe pain and/or limited function, the people most in need of surgery.
- Some risk variables that might be traditionally considered as predictors of worse outcomes are positively associated with achieving a SCB because patients with more severe symptoms at baseline have more opportunity for improvement. Patients on our TEP and on our Patient Working Group supported this improvement threshold.
- The TEP supported a lower opportunity for patients with high scores preoperatively, indicating that mild symptoms, to reach substantial clinical benefit improvement. TEP members were in favor of a measure that dis-incentivized inappropriate surgery and clinicians performing major elective surgery on patients will little opportunity for

[8] Lyman S and Lee YY. What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? *Clinical Orthopaedics and Related Research®.* 2018;467(12):2432-2441.

benefit.

- **Issue 13:** An NQF Panel member voiced concerns with the lack of adjustment for non-English speakers, given that the KOOS, Jr. and HOOS, Jr. are only offered in English.
    - **Developer Response 13:** We do not have available to us a variable representing primary or spoken language, preventing risk adjustment consideration. Active efforts to make the HOOS, JR and KOOS, JR available in additional languages are ongoing. We will recommend to CMS that it considers collecting preferred language status for possible risk adjustment.
- **Issue 14:** An NQF Panel member noted concern that the only social risk factor included is health literacy.
    - **Developer Response 14:** Health literacy is a potent predictor and associated with a range of social risk factors. In addition, as noted in the NQF Testing Attachment, Section 2b3.4b, the results of the social risk factor testing did not provide evidence of significant differences in measure results. However, we did find that social risk factors were significantly associated with response and therefore, we included social risk in our non-response adjustment of the measure. As this measure assesses patients undergoing an elective procedure where known disparities exist, we will recommend CMS continues to assess the impact of social risk for this measure over time.
- **Issue 15:** An NQF Panel member noted large differences between NQF #1550 groups on data element, and that few patients appear to report substantial clinical improvement, noting that this could be because the bar is set too high, or ceiling effects of the measures, or both.
    - **Developer Response 15:** It appears that this Panel member is referring to Figure 1 in the NQF Testing Attachment when referring to NQF#1550 (the THA/TKA Complications measure used for Empiric Measure Score validity assessment). The comment regarding “few patients appear to report substantial clinical improvement” is not understood. This figure shows that hospitals with worse than the national average complication rates have a median RSIR just above 50%, whereas hospitals at the national average complication rates have a median RSIR near 65% and hospitals with better than national average complications rates has a median RSIR at approximately 70%. Table 14 of the NQF Testing Attachment notes that the risk-standardized mean RSIR for hospitals is 60%.
- **Issue 16:** An NQF Panel member noted that the THA/TKA PRO-PM RSIR with the hospital risk standardized complication rate (NQF: 1550) displayed box plots with evidence of considerable validity in results at the mean. A plot of the association of pass/fail on each measure at the hospital level would have been helpful.
    - **Developer Response 16:** As outcome measures are often reported as point estimates with uncertainty ranges reflecting the statistical uncertainty inherent in outcome measurement, we felt it better to represent the statistical uncertainty that CMS reports for NQF 1550 than to report validity using only the point estimate and without acknowledging the statistical uncertainty. As CMS has not yet indicated it plans for reporting RSIRs, we did not apply any calculation of statistical uncertainty to the RSIRs.

Other General Comments

- **Issue 1:** NQF Panel Member #1 state that specifications of the measure identified that it was both risk-stratified and risk-adjusted, and that the difference between these two terms were not provided.
    - **Developer Response 1:** This statement is not consistent with the information we provided. The measure is identified in different sections as risk-standardized (or risk-adjusted) but not risk-stratified. The terms risk-standardized and risk-adjusted both signify that the measure is risk-adjusted.
- **Issue 2:** NQF Panel Member #8 noted that the variables listed with measure specifications in S5 for risk adjustment did not match those listed in 2b3.
    - **Developer Response 2:** In S5, we identify the data elements listed as those used to define the numerator and for risk adjustment that are collected with PROM data; this is not intended to be a complete list of risk variables in the risk adjustment model.
- **Issue 3:** There was a request for explanation of why multiple risk adjustment variables in the risk-adjustment model were included that where not significant?
    - **Developer Response 3:** When building claims-based models, we have previously used the strength of association between the risk variable and the measure outcome to empirically guide risk variable selection. When expert input deems it appropriate, we force in additional risk variables, such as those that indicate frailty, that might have an important influence on the measure outcome and yet might not be selected for the model based purely on statistical considerations. In this way, our risk models always reflect both empirical data and clinical input. This approach has produced robust risk models that have been repeatedly and successfully validated against medical record data. For this measure, we applied the same principles, but recognize that PRO-PM development, particularly that based upon a voluntary data sample, may require a greater reliance on clinical input to select risk variables than traditional claims-based outcome measures. Therefore, for this measure, we conducted analyses to evaluate two approaches to risk model development for each PROM outcome – one used a purely data-driven approach (referred to as the empirically derived model) and another used candidate risk variable selection based on empirical findings in the literature, review of data-driven risk factors, and iterative TEP and clinical expert input and ranking of importance and feasibility of risk variables for a THA/ TKA PRO-PM (referred to as a clinically derived model). We identified an extensive list of risk variables for consideration in the development of the risk model(s), through a systematic literature review and environmental scan, as well as from orthopedists surveyed about what risk variables they consider important in predicting THA/TKA outcomes. In consultation with the Technical Working Group and the TEP and through detailed public comments from specialty societies, we focused on candidate risk-adjustment variables of interest that were clinically relevant and had an evidence-based relationship with clinical outcomes following elective primary THA or TKA. Likewise, we considered several potential data sources, including administrative claims, registry- or clinician-provided data, and patient-reported sources. In addition to clinical risk variables that have been collected de novo and evaluated for inclusion in the final measure risk model, all diagnostic codes from administrative claims during the 12 months prior to the THA/TKA procedure or secondary diagnosis codes during the index admission except those associated with

potential complications during the index admission were evaluated for possible inclusion in the risk model. Recognizing the thorough vetting of risk variables for this risk model, we determined to keep variables in the model that may not reach statistical significance in our data with an understanding that our sample may be limited and that this risk model will be implemented more broadly.

- **Issue 4**: An NQF Panel member suggested that the creation of the HOOS, JR and KOOS, JR instruments were never discussed, and that they would have liked evidence of the content coverage (content validity) for each measure.
    - **Developer Response 4:** The HOOS, JR and KOOS, JR were developed at the Hospital for Special Surgery by Stephen Lyman and colleagues. The instruments are non-proprietary, free to use, and were validated in 2016. Reliability and validity testing conducted for these PROM surveys is reported in manuscripts on the validation of these instruments[9],[10] and noted in the NQF Testing Attachment form, Sections 2a2.2 and 2a3.3.

[9] Lyman S, et al. Validation of the HOOS, JR (see footnote 1, page 2)

[10] Lyman S, et al. Validation of the KOOS, JR (see footnote 2, page 2)