NATIONAL QUALITY FORUM
Driving measurable health improvements together

Scientific Methods Panel Discussion Guide

FALL 2021 EVALUATION CYCLE
October 26-27, 2021

This report is funded by the Centers for Medicare & Medicaid Services under contract HHSM-500-2017-00060I - HHSM-500-T0001.

Contents

Scientific Methods Panel Discussion Guide 1
Contents 2
Background 3
Measures for Discussion (Brief) 4
Subgroup 1 4
Subgroup 2 4
Measures That Passed (Not Pulled for Discussion) (Brief) 4
Subgroup 2 4
Measures For Discussion (Detailed) 5
Subgroup 1 5
Subgroup 2 19
Appendix A: Measures That Passed (Not Pulled for Discussion) (Detailed) 23
Subgroup 2 23
Appendix B: Additional Information Submitted by Developers for Consideration 42
Subgroup 1 42
Subgroup 2 84

Background

The Scientific Methods Panel (SMP) provides National Quality Forum (NQF) Standing Committees with evaluations of submitted complex measures' Scientific Acceptability (specifically, the "must-pass" subcriteria of reliability and validity), using NQF's standard measure evaluation criteria for new and maintenance measures.

This discussion guide contains details of the complex measures submitted for evaluation during the fall 2021 measure evaluation cycle. It also contains summaries of preliminary measure analyses and responses to these analyses composed by developers. The SMP utilizes this document during measure evaluation meetings to facilitate conversations between the Panel, measure developers, and NQF staff. Measures are slated for discussion and revote at the SMP's measure evaluation meeting if consensus was not reached during the preliminary review, or if a measure did not pass a sub-criterion and the developer organization provided a written response to the SMP's comments. Additionally, SMP members and NQF staff may pull a measure for further discussion as they see fit, even if the measure passed preliminary review. This cycle, 12 complex measures were evaluated by the SMP. Seven are up for discussion and revote. Four of the seven have been pulled by SMP members or NQF staff for further discussion, although they have passed NQF's Scientific Acceptability criterion.

Following the SMP's review of the complex measures, those that pass Scientific Acceptability move on to their respective Standing Committees for measure evaluation of the remaining NQF standard measure evaluation criteria (specifically, Importance to Measure and Report, Feasibility, Usability and Use, and requirements for Related and Competing Measures). Measures that do not pass the SMP may be pulled by a Standing Committee member for further discussion and revote if it is an eligible measure. A measure is eligible for revote if the SMP found none of the following:

- Inappropriate methodology or testing approach applied to demonstrate reliability or validity
- Incorrect calculations or formulas used for testing
- Description of testing approach, results, or data is insufficient for SMP to apply the scientific acceptability sub-criteria
- Appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements

Please refer to "Scientific Methods Panel: Frequently Asked Questions" in NQF's standard measure evaluation criteria for further details on this process.

Measures for Discussion (Brief)

Subgroup 1

- 3649e Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) electronic clinical quality measure (eCQM) (Brigham and Women's Hospital)
    - Reliability: H-1; M-3; L-6; I-0 CNR
    - Validity: H-0; M-8; L-1; I-1 Pass
- 3650e Risk-standardized inpatient respiratory depression (IRD) rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM (Brigham and Women's Hospital)
    - Reliability: H-0; M-7; L-4; I-0 Pass
    - Validity: H-0; M-3; L-8; I-0 No Pass
- 3652e Risk-standardized prolonged opioid prescribing rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM (Brigham and Women's Hospital)
    - Reliability: H-0; M-7; L-3; I-1 Pass
    - Validity: H-2; M-5; L-4; I-0 Pass
- 3638 Care Goal Achievement Following a Total Hip Arthroplasty (THA) or Total Knee Arthroplasty (TKA) (Brigham and Women's Hospital)
    - Reliability: H-0; M-1; L-5; I-3 No Pass
    - Validity: H-0; M-3; L-3; I-3 No Pass
- 3639 Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM) (Yale CORE/Centers for Medicare & Medicaid Services)
    - Reliability: H-3; M-3; L-1; I-2 Pass
    - Validity: H-0; M-7; L-1; I-1 Pass
- 3667 Days at Home for Patients with Complex, Chronic Conditions (Yale CORE/Centers for Medicare & Medicaid Services)
    - Reliability: H-5; M-6; L-0; I-0 Pass
    - Validity: H-2; M-7; L-2; I-0 Pass

Subgroup 2

- 0689 Percent of Residents Who Lose Too Much Weight (Long-Stay) (Centers for Medicare & Medicaid Services)
    - Reliability: H-3; M-5; L-3; I-0 Pass
    - Validity: H-1; M-5; L-5; I-0 CNR

Measures That Passed (Not Pulled for Discussion) (Brief)

Subgroup 2

- 3633e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Level) (UCSF/Alara Imaging)
    - Reliability: H-9; M-2; L-0; I-0 Pass
    - Validity: H-5; M-6; L-0; I-0 Pass
- 3662e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinical Group Level) (UCSF/Alara Imaging)

  - Reliability: H-8; M-3; L-0; I-0 Pass
  - Validity: H-7; M-4; L-0; I-0 Pass
- 3663e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level) (UCSF/Alara Imaging)
  - Reliability: H-9; M-2; L-0; I-0 Pass
  - Validity: H-6; M-5; L-0; I-0 Pass
- 3665 Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood (American Academy of Hospice and Palliative Medicine)
  - Reliability: H-3; M-6; L-1; I-1 Pass
  - Validity: H-3; M-5; L-3; I-0 Pass
- 3666 Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain (American Academy of Hospice and Palliative Medicine)
  - Reliability: H-4; M-5; L-2; I-0 Pass
  - Validity: H-2; M-6; L-3; I-0 Pass

Measures For Discussion (Detailed)

Subgroup 1

Measure# 3649e: Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) electronic clinical quality measure (eCQM) (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This measure quantifies the risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) at the clinician group level for adults 18 years and older across all payers. The rate is expressed as a percentage where a lower rate is indicative of higher quality care. The outcome is defined as any of the specified complications occurring from the date of index admission to 90 days following discharge (or procedure encounter if the procedure is done on an outpatient basis)
- **Type of measure:** Outcome
- **Data source:** Electronic Health Record
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 10 risk factors
- **Sampling allowed**: None
  - **Ratings for reliability:** 1 high 3 moderate 6 low and 0 insufficient → Consensus not reached
  - Reliability testing conducted at the Patient or Encounter level:
    - The developer used the Feasibility Scorecard to assess electronic health record (EHR) data availability, accuracy, terminology standards, and workflow.
    - All 23 data elements scored a 1/1 on the Feasibility Scorecard for both Cerner and EPIC sites.
    - One SMP member questions how the patient demographic data "sheds light" on element-level reliability.
  - Reliability testing conducted at the Accountable Entity level:

- The developer used the test-retest approach to test the reliability of the predicted/expected ratios at the clinician group level, comparing the agreement across clinician-groups on the performance measure.
  - Predicted/expected ratios for the six clinician groups (i.e., six EPIC sites, 11 Cerner sites) ranged from 0.719-1.404 with 95 percent confidence intervals (CI). Adjusted rate overall = 3.66. Spearman rank correlation= 0.978.
- For variability across clinician-groups, the intraclass correlation coefficient (ICC) = 0.006 (95 percent CI: -0.017-0.027). The developer mentions that the low ICC is likely due to the small clinician-group sample sizes and little variation in performance across groups.
- Assessment of the risk adjustment's logistic regression model was assessed using Hosmer-Lemeshow calibration, p=0.820 (test = 0.795, validation = 0.846).
- To determine the strength of the model in predicting complication events, the developers calculated the C-statistic = 0.672 (test = 0.674, validation = 0.670).
- One SMP member raised concern with the generalizability of the results for 17 providers during measure score level testing.
- One member raised questions about the measurement's performance period for clinician-groups who will use and report this measure.
- One SMP member raised concern that the four years of data used for testing and validation may contain higher sample sizes in each split half sample then that available when actually implemented for a single calendar year. The SMP member is concerned that the Spearman correlation coefficient results for this testing may be overestimated (i.e., across two calendar years instead of one).
- One SMP member raised concern with the ICC's wide confidence interval (CI) around the ICC estimates and that different versions of the ICC estimates are presented in the testing results.
- One SMP member raised concern with the low ICC (0.006) and the generalizability of the results.
- One SMP member raised concern with the statistical uncertainty as the results show high correlation between point estimates across training and validation samples.

- **Ratings for validity:** 0 high 8 moderate 1 low and 1 insufficient → Measure passes with MODERATE rating
- Validity testing conducted at the Patient or Encounter level:
  - The developer conducted reliability testing at the data element level. A sample of 217 EPIC patients and 25 Cerner patients were analyzed by conducting manual chart abstraction and comparing numerator, denominator, and exclusion data to the eCQM

calculation, noting any disagreements between the EHR and the eCQM. Kappa statistics were used to assess agreement (0.60-0.90 indicates strong agreement, above 0.90 indicated perfect agreement).
    - Kappa coefficients for agreement ranged from 0.8333 to 0.9495 with agreement percentages ranging between 88.89 percent to 96.67 percent for EPIC sites. Kappa results were 84 percent in the Cerner site testing.
  - One SMP member raised concern with the provided complication rates being at the hospital level and not at the clinician-group level.
- Validity testing conducted at the Accountable Entity level:
  - The developers assessed face validity through a seven-member Technical Expert Panel (TEP) process in which the TEP was engaged throughout the electronic measurement development process, providing feedback during certain points to the developer. Final measure specifications were presented to the TEP for review, in addition to stakeholder feedback and developer rationale as to the meaningfulness of the measure. The TEP provided a silent vote through an online poll platform.
    - 85 percent (6/7) of TEP members agreed that this measure was an accurate reflection of quality and could be used to distinguish good from poor clinician-group level care quality related to patient safety.
  - For empirical validity, the developer mentions that the eCQM's data elements are harmonized with NQF #1550. The developer for #1550 completed a validation study among multiple national sites to demonstrate the correlation between claims data used to code complications and data documented within the EHRs. The study demonstrated strong agreement between patients undergoing a primary THA or TKA and experiencing complications in both claims based and electronic medical record data.
  - Some SMP members raised concern with measure score validity testing, noting that only face validity was assessed. Additionally, some SMP members mentioned that the TEP was asked only one question during face validity testing.
  - A few members expressed concerns with low value of the c-statistic and the small variation and implications for assessing meaningful differences in performance.
  - One SMP member raised concerns with the risk adjustment model, noting concerns with the methods for selection of risk factors.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action Items:**
  - The structure, testing methods, and results, as well as SMP expressed concerns, are similar to measures 3650e currently under review by the SMP.

- Considering the low testing sample, is generalizability of the 17 providers/EHR sites in the accountable entity level reliability sample demonstrated?
- Does the split-half four years (i.e., two two-year samples) of testing data versus a single calendar year or the hospital versus clinical groups complications rates effect attribution? Do the wide CI around the ICC estimates and different estimates effect demonstrate ICC concerns?
- With the variety of reliability and validity testing conducted, and with reliability was rated lower than validity, does the SMP have concerns with the appropriateness of the reliability testing (e.g., some members commented that the comparison of socio-demographic characteristics did not seem to shed light on reliability) and their perspective on the results?
- How does the SMP view the validity and methods for building the risk adjustment model?

Measure# 3650e: Risk-standardized inpatient respiratory depression (IRD) rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM (Pulled by an SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This eCQM estimates the risk-standardized inpatient respiratory depression (IRD) rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) at the clinician group level for adults 18 years and older across all payers.
- **Type of measure:** Outcome
- **Data source:** Electronic Health Records
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with 10 factors
- **Sampling allowed**: None
    - **Ratings for reliability:** 0 high 7 moderate 4 low and 0 insufficient → Measure passes with MODERATE rating
        - Reliability test sample. The developer used 17 total orthopedic groups; six from MGB and 11 from a CERNER site. All provider groups use the CERNER EHR system and perform between 25 and 1200 THA/TKA surgeries per year. Two orthopedic groups were excluded from the measure since they performed <25 surgeries during the measurement period.
    - Reliability testing conducted at the Patient or Encounter level:
        - The developer performed a review of low volume of 30 random patients to evaluate the accuracy of eCQM abstraction.
        - All data elements abstracted by the eCQM matched with the information within the EHR.
        - The developer also compared the sociodemographic characteristics of patients included in the test to validation samples and found no differences between sites or clinician groups.
        - SMP members raised questions on whether this method of reliability testing of sociodemographic characteristics across two subgroups is sufficient to demonstrate reliability.

- Reliability testing conducted at the Accountable Entity level:
    - The developer performed reliability testing at the accountable entity-level using a test-retest approach to examine the reliability of the predicted/expected ratios at the clinician group level.
    - The developer found that the test and validation samples gave similar ranking of the 17 clinician groups with respect to the predicted/expected ratios with a Spearman rank Correlation of 0.767 between the two samples. The developer also estimated a low ICC between clinician groups and the ICC value at 0.069151.
    - SMP members noted the high correlation statistic but raised concerns that that the ICC presented by the developer raises reliability concerns.
    - The four years of data used for testing and validation may contain higher sample sizes in each split half sample then that available when actually implemented for a single calendar year. The SMP member is concerned that the Spearman correlation coefficient results for this testing may be overestimated (i.e., across two calendar years instead of one).
    - One SMP member raised concern with the ICC's wide CI around the ICC estimates and the different versions of the ICC estimates presented in the testing results.

- **Ratings for validity:** 0 high 3 moderate 8 low and 2 insufficient → Measure does not pass with LOW rating
- Validity testing conducted at the Patient or Encounter level:
    - The developer assessed the frequency of data elements needed for risk adjustment and data element agreement between manual chart review and EHR calculation.
- Validity testing conducted at the Accountable Entity level:
    - The developer convened a TEP to assess the face validity of the measure. The developer reported that 3/7 (42.86 percent) TEP members agreed that the measure was actionable to improve quality of care.
    - SMP members noted that face validity testing was conducted by the developer with rather low results.
    - The developers risk adjusted both the predicted and expected numerator events for age, gender, type of surgery (THA/TKA), insurance, race, household income, English as primary language, smoking status, body mass index and comorbidities.
    - The developer provided Hosmer-Lemeshow Calibration p-values for the risk adjustment model. For the Test Sample, p=0.55599, and for the Validation sample, p=0.98401.
    - Several SMP members raised concerns on the conceptual rationale for the risk adjustment strategy. SMP members noted that the use of social risk factors: race, income, insurance status should not be

used in this measure without a strong conceptual framework for why these might influence IRD.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
    - The structure, testing methods and results, as well as SMP expressed concerns are similar to measures 3649e currently under review by the SMP.
    - Does the SMP have concerns related to the small testing samples for determining reliability and validity, and risk adjustment, as well as determining meaningful differences within and between populations (e.g., vulnerable populations)?
    - For reliability, how does the SMP interpret the low Spearman correlation coefficient in light of the ICC score?
    - How does the SMP interpret the low TEP (3/7, 43 percent) face validity testing results?

Measure# 3652e: Risk-standardized prolonged opioid prescribing rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This eCQM assesses the percentage of patients 18 years and older across all payers who were not previously exposed to opioids within 90 days prior to the THA/TKA procedure and who were prescribed opioids for 42 days (six weeks) following an elective primary THA/TKA.
- **Type of measure:** Process
- **Data source:** Electronic Health Records
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with eight risk factors
- **Sampling allowed**: None
    - **Ratings for reliability:** 0 high 7 moderate 3 low and 1 insufficient → Measure passes with MODERATE rating
        - For the specifications, an SMP member raised a question as to the 42-day interval selected for the measure.
        - SMP members were concerned about whether the measure could be generalizable to EHRs outside of EPIC and Cerner. The SMP expressed concern the limit of the EHR testing would be generalizable to other EHR systems.
        - Several SMP members questioned whether the measure is appropriately categorized as an outcome measure, rather than a process measure.
    - Reliability testing conducted at the Patient or Encounter level:
        - The developer noted that the NQF eCQM Feasibility Scorecard was used to assess the EHR data availability, accuracy, terminology standards, and workflow. The measure scored a 1/1 for all 22 data elements for availability, accuracy, data standards, and workflow within both EHR systems used in testing (i.e., EPIC, Cerner).
        - The developer compared sociodemographic characteristics of patients included in test and validation samples and found there

were no differences at the patient level (p = 0.12-0.68) or between clinician group (p = 0.99).
    - SMP members questioned whether comparing sociodemographic factors across test and validation samples adequately demonstrates reliability.
- Reliability testing conducted at the Accountable Entity level:
    - The developer used a test-retest approach to assess the reliability of the predicted/expected ratios at the clinician group level.
    - The developer estimated how the two random samples agree using a Spearman correlation coefficient. The measure had a spearman rank correlation of 0.8182 for THA procedures and 0.8909 for TKA procedures.
    - An ICC was conducted to assess variability across clinician groups. The ICC for THA was 0.0929 (95 percent CI = -0.043-0.197) and the ICC value for TKA was 0.11675 (95 percent CI = -0.013-0.217).
    - SMP members raised concerns with the ICC results and whether the measure is able to meaningfully capture variation in provider performance.
    - The four years of data used for testing and validation may contain higher sample sizes in each split half sample then that available when actually implemented for a single calendar year. The SMP member is concerned that the Spearman correlation coefficient results for this testing may be over-estimated (i.e., across two calendar years instead of one).
    - One SMP member raised concern with the ICC's wide CI around the ICC estimates and the different versions of the ICC estimates presented in the testing results.
- **Ratings for validity:** 2 high 5 moderate 4 low and 0 insufficient → Measure passes with MODERATE rating
- Validity testing conducted at the Patient or Encounter level:
    - The developer found that the manual chart review and the eCQM had perfect agreement (final Kappa = 1), and Cerner site chart reviews surpassed the minimum threshold for validity (70 percent agreement).
    - Some SMP members expressed concern comments on missing data here and potential lack of ability to generalize to other EHRs.
    - One SMP member expressed concern with the small size of the TEP and that the assessment was based on a single question.
- Validity testing conducted at the Accountable Entity level:
    - The developer tested the face validity by using a panel of experts, 7/7 of whom agreed that the measure is an accurate reflection of quality, and that it can be used to distinguish between good and poor quality.
    - One SMP member expressed concerns about the multiple rounds of data testing required to reach validity, suggesting that the process would need to be repeated with each different EHR.

- One SMP member noted that there are no data elements missing in Epic but "days supplied" is missing about 34 percent of the time in Cerner. The developer adjusted the predicted and expected extended use rates for age, sex, race, household income, English as primary language, body mass index, and comorbidities.
- A Hosmer-Lemeshow calibration was performed to assess the goodness of fit for the logistic regression model with risk adjustment with P values of 0.618 (THA), and 0.643 (TKA).
- A C-statistic was used to assess the strength of the model in predicting prolonged prescribing events; the C-statistic for the risk model is 0.708 for hip and 0.655 for knee.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
  - How does the SMP interpret the Spearman Correlation Coefficient and ICC, in terms of demonstrating accountable entity reliability?
  - Does the expanded timeline for the test-retest sample effect the reliability of the measure?
  - How does the SMP interpret the validity testing results?
  - How does the more than one-third Cerner missing data effect validity?
  - Is the risk adjustment strategy appropriate for this process measure?

Measure# 3638: Care Goal Achievement Following a Total Hip Arthroplasty (THA) or Total Knee Arthroplasty (TKA)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The percentage of adult patients 18 years and older who had an elective primary total hip arthroplasty (THA) or total knee arthroplasty (TKA) during the performance period AND who completed both a pre- and post-surgical care goal achievement survey and demonstrated that 75 percent or more of the patient's expectations from surgery were met or exceeded.
  The pre- and post-surgical surveys assess the patient's main goals and expectations (i.e., pain, physical function and quality of life) before surgery and the degree to which the expectations were met or exceeded after surgery.
  The pre-surgical data collection timeframe will be zero to 90 days before surgery and the post-surgical data collection timeframe will be 90 to 180 days after surgery.
  The patient-reported outcome-based performance measure (PRO-PM) score is derived by calculating the differences between the pre-surgical and the post-surgical surveys. A higher score indicates greater care goal achievement. The measure will be reported as two risk-adjusted rates stratified by THA and TKA.
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Registry Data; Claims; Electronic Health Records; Instrument-Based Data; Paper Medical Records
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical Risk Model with three factors
- **Sampling allowed**: None

- **Ratings for reliability:** 0 high 1 moderate 5 low and 3 insufficient → Measure does not pass with LOW rating
- Reliability testing conducted at the Patient or Encounter level:
    - The developer tested interrater reliability through chart review.
        - Data was obtained from an Electronic Data Warehouse (EDW) and through manual chart review (n=68; 34 THA and 34 TKA patients).
        - Alignment between the manual reviewers and the EDW overall was 97.1 percent agreement (kappa value of 0.93).
        - Alignment between manual reviewers and EDW data elements for THA and TKA had 100 percent (kappa value was 1.00) and 94.1 percent agreement (kappa value of 0.87), respectively.
        - The overall agreement between the reviewers and the electronic data warehouse ranged from 89.9-99.2 percent.
        - One SMP member raised concern with the threshold of 0.7 set for acceptable internal consistency reliability.
- Reliability testing conducted at the Accountable Entity level:
    - The developer performed reliability testing of the accountable entity (i.e., measure score) using a signal-to-noise ratio (SNR) approach.
        - The SNR generated by the developer was 0.00118 for THA and 0.00004 for TKA.
    - Some SMP members raised concern with clinician-group reliability, noting the low ICC estimates and small sample sizes.
    - One member acknowledged that the developer mentions the effect of low sample size on the ICC estimate; however, the SMP member raised concern with the lack of between-practice variation and the reliability of this measure at the practice level.
    - Some SMP members mentioned that reliability testing is sufficient at the patient or encounter (i.e., data element) level, yet inadequate at the clinician-group practice level due to small sample size, low variability of scores across practices, and no assessment of nonresponse bias.
- **Ratings for validity:** 0 high 3 moderate 3 low and 3 insufficient → Measure does not pass with LOW/INSUFFICIENT rating
- Validity testing conducted at the Accountable Entity level:
    - Face validity was assessed from a six-person TEP, which was convened to provide input on the conditions, groupings, and modeling. Public commenting was also requested.
        - The majority of TEP members agreed that the measure had suitable face validity.
    - Empirical validity was assessed through measure known-groups and measure discriminant testing.
        - Measure known-groups validity was tested through a one-question post-surgical satisfaction survey. The developer

did not calculate the Pearson correlation due to small sample size.
    - Measure discriminant validity was tested by comparing the means of care goal achievement (CGA) PRO-PM results by joint for clinician-groups with a minimum case-volume requirement of at least 25 patients. The THA adjusted mean was 58.4 percent (*SD*= 11.6 percent) and the TKA adjusted mean was 41.3 percent (*SD*=6.3 percent).
- Several SMP members raised various concerns with the empirical validity testing and interpretation due to the small sample sizes overall and for the risk adjustment model, testing methodology, apparent homogenous populations, lack of population variability (including social risks), and inconclusive results during measure known-groups testing.
- One SMP member raised concerns whether the risk adjustment model adequately balances priori decisions for variable inclusion with metrics of fit after model testing. No evidence for the validation of the risk adjustment model is present.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
    - Is there any new information that the SMP would like to revisit related to the reliability vote, specifically expressed concerns that reliability of the PROM may be biasing the judgement of reliability of the PRO-PM?
    - The SMP should discuss the effect that small sample sizes have on measure's reliability and validity.
    - How does the SMP interpret the validity testing results based on the sample size, generalizability concerns, lack of potentially relevant social risks impacts, and risk factor selection in the adjustment model?

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

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This patient-reported outcome-based performance measure uses the same measure specifications as the NQF-endorsed (NQF # 3559) hospital-level risk-standardized improvement rate (RSIR) following elective primary THA/TKA with the following exception: this measure attributes the outcome to a clinician or clinician group. Specifically, this measure will estimate a clinician-level and/or a clinician group-level RSIR following elective primary THA/TKA for Medicare fee-for-service (FFS) patients 65 years of age and older. Improvement will be calculated with patient-reported outcome data collected prior to and following the elective procedure. The preoperative data collection timeframe will be 90 to 0 days before surgery and the postoperative data collection timeframe will be 270 to 365 days following surgery.
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Claims; Instrument-Based Data

- **Level of analysis:** Clinician: Group/Practice; Clinician: Individual
- **Risk-adjusted:** Statistical risk model with 19 factors
- **Sampling allowed**: None
    - **Ratings for reliability:** 3 high 3 moderate 1 low and 2 insufficient → Measure passes with MODERATE rating
    - Reliability testing conducted at the Patient or Encounter level:
        - The developer did not conduct patient- or encounter-level testing of the PRO-PM in the specified measure population, timeframe, and setting as required. Rather, they cited used test-retest and internal consistency to assess reliability of both PRO-PM instruments or PROMs (i.e., HOOS, JR and KOOS, JR). Internal consistency was calculated using the Pearson Separation Index (PSI) for both instruments. Internal consistency ranged from 0.84-0.87.
        - Intra-class correlations for reliability were between four dimensions (Pain, Symptoms, Activities of Daily Living, Sport and Recreation Function, and Quality of Life) of the HOOS, JR and the KOOS, JR with ranges from 0.75 to 0.97.
    - Reliability testing conducted at the Accountable Entity level:
        - The developer performed reliability testing at the measure score-level using a signal to noise ratio (SNR) approach. Among clinicians and clinician-groups with five and 10 cases, the SNR yielded median reliability scores ranging from 0.70-0.79 and 0.79-0.85, respectively. The mean reliability score was 0.69 (SD 0.16) for clinicians with at least five cases.
        - Among clinicians and clinician-groups with at least 25 cases, the SNR ratio yielded median reliability scores ranging from 0.79-0.97 (median 0.87, interquartile range [IQR] 0.09) and 0.79-0.99 (median 0.92, IQR 0.10), respectively.
        - One SMP member raised concern in regard to variation in responses as it relates to social risk (i.e., race) and that the experiences among racial groups may be underrepresented in the sample.
    - **Ratings for validity:** 0 high 7 moderate 1 low and 1 insufficient → Measure passes with MODERATE rating
    - Validity testing conducted at the Patient or Encounter level:
        - The developer evaluated responsiveness for both instruments using standardized response means and then compared against two other previously validated PROMs.
            - External validity was evaluated for both instruments using Spearman's correlation.
            - a) Correlations ranged from 0.84- 0.94 for HOOS, JR testing.
            - b) Correlations ranged from 0.72- 0.91 for KOOS, JR testing.
        - The floor and ceiling effects for HOOS, JR were (0.6 percent – 1.9 percent) and (37 percent – 46 percent), respectively.
        - The floor and ceiling effects for KOOS, JR were (0.4 percent – 1.2 percent) and (18.8 percent – 21.8 percent), respectively.

- One SMP member raised concern with the potential for measurable improvement related to the floor and ceiling effect. For the HOOS, JR testing, the ceiling effect (37 percent to 46 percent) did not meet the 22 points to support substantial clinical benefit.

- Validity testing conducted at the Accountable Entity level:
  - Face validity was assessed by asking a 17-member TEP to respond to two statements using a six-point scale.
  - 76 percent either strongly or moderately agreed with the statement that this measure, as specified, will provide a valid assessment of improvement in functional status and pain following elective, primary THA/TKA. Fifty-three percent either strongly or moderately agreed with the statement that this measure, as specified, can be used to distinguish between better and worse quality care among clinicians and clinician groups.
  - Some SMP members expressed interest in observing descriptive characteristics for those patients with no response to allow for construction of models to adjust for nonresponse prior to assessing reliability.
  - Several SMP members raised concern with non-response bias and the accuracy of the developer's validity assessment as 37 percent of the sample was excluded due to missing PRO scores, 10 percent due to missing risk factors, and 2 percent without clinician attribution.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developer
- **Action items:**
  - Is the measure reliable without the PRO-PM required patient or encounter testing of the measure "providing 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?"
  - Does the testing sample adequately represent diverse populations to be generalized among and between all groups?
  - How do the large non-response bias volumes and missing risk factors, especially in social risk populations, impact validity?

Measure# 3667: Days at Home for Patients with Complex, Chronic Conditions (Pulled by SMP Member)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This is a provider group-level measure of days at home or in community settings (that is, not in acute care such as inpatient hospital or emergent care settings or post-acute settings such as Skilled Nursing Facilities (SNFs)) among adult (age 18 years or older) Medicare FFS beneficiaries with complex, chronic conditions who are aligned to participating provider groups. The measure includes risk adjustment for differences in patient mix across provider groups, with an adjustment based on patients' risk of death. An additional adjustment that accounts for patients' risk of transitioning to a long-term nursing home is also applied to encourage home- and community-based care in alignment with Centers for Medicare &

Medicaid Services' (CMS) policy goals. A higher risk-adjusted score indicates better performance.

- **Type of measure:** Outcome
- **Data source:** Claims
- **Level of analysis:** Accountable Care Organization
- **Risk-adjusted:** Statistical risk model with 52 risk factors
- **Sampling allowed**: None
    - **Ratings for reliability:** 5 high 6 moderate 0 low and 0 insufficient → Measure passes with MODERATE rating
    - Reliability Specifications
        - Many SMP members found the measure's specifications confusing and occasionally arbitrary. They report that it did not appear to align throughout the measure submission and choices made were not clear. This was especially true of the denominator statement, which lacked the target population, conditions, settings, etc.
        - The SMP expressed concerns that several concepts included in the submission were not documented as exclusions in the specification. Their absence threatens the measure's validity and may incentivize under-treatment of conditions potentially outside the control of the accountable entity. The SMP also questioned whether the developer considered other exclusions related to specific reasons for being accepted into acute care/ED that might not indicate low quality of the accountable entity. Should patients be excluded from the numerator and denominator days after death occurs?
        - An SMP member expressed concerns with adjusting for transitions to the nursing home, which purports that moving from home to a nursing home is always negative. Permanent nursing home admissions requiring skilled nursing care also incorporates the patient's available personal and community resources, which is also not at the control of the accountable entity.
        - SMP members also pointed out that the unit of analysis vacillates between accountable care organization (ACO) and provider group. An ACO and medical group aren't the same unit of analysis.
        - One SMP member questioned whether this measure that combines multiple risk models into a single overall score should be considered a cost composite measure.
    - Reliability testing conducted at the Accountable Entity level:
        - The developer tested the measure using a split half methodology, using data from 2017-2018. They reported an ICC of 0.8326 for the final Days at Home outcome metric between the two samples. Beyond a "split-half" analysis, the form of ICC is not described.
        - SMP members noted that the use of split-half methodology is better suited for federal accountability programs with multiple years of data, particularly because ACO assignment rules are adjusted annually. Further, the reported ICC may under-estimate true reliability because scores are estimated using only half of each provider's data.

- **Ratings for validity:** 2 high 7 moderate 2 low and 0 insufficient → Measure passes with MODERATE rating
- Validity testing conducted at the accountable entity level:
    - Construct validity was assessed using Pearson correlations with six other ACO-level measures (i.e., ACO-8, ACO-38, ACO-13, ACO-43, ACO-35, and ACO-1) of quality conceptually related to excess days of care for patients with complex chronic diseases.
    - Pearson's correlations ranged between -0.549 and +0.048 resulting in a high inverse for unplanned admissions (expected), moderate with other measures, no correlation with fall risk, and an unexpected inverse correlation with patient experience.
    - This was attributed to the range of focus for the measures compared. The latter was assessed on a smaller sample size and not risk adjusted for clinical variables but warrants further explanation.
    - The developer performed face validity testing of the measure specifications and the appropriateness for quality assessment at the ACO level with 19 of 21 responding TEP members. For the statement posed to the TEP, "The Days at Home measure, as specified, can be used to distinguish between better or worse performance at ACOs or provider groups," two members indicated "strongly agree," 15 indicated "agree," and two indicated "somewhat agree."
    - For each patient, adjusted days at home calculates "excess days in care" with three risk models: 1) risk-adjusted days in acute care settings or SNFs among days alive in the year, 2) risk of mortality, and 3) risk of transition to nursing home. "Excess days in care" are updated based on risk of death and risk of transition to nursing home care and then averaged across each provider group to produce the final measure scores.
    - Some SMP members noted that there are three different risk adjustment models used and expressed concerns about lack of clarity about whether/how they were combined to get a single score and the validity of the approach.
    - One SMP member indicated it is not clear why primary death data was not used, but a death risk model was used instead.
    - SMP members had concerns with the risk adjustment methodology, testing, and results. Some members noted that the measure construction approach inappropriately lacked adjustment for many variables without theoretical or empirical justifications, as well as a potential arbitrary measure weighting.
    - Some SMP members expressed concerns related to multiple data elements/concepts not considered/included in the model.
    - Several SMP members commented that many decisions regarding social risk factors appeared arbitrary and were not persuasive.
    - The c-statistic was 0.738 for the mortality model, 0.760 for nursing home transition. Deviance from R-squared was 0.170 for the days in

care model. Spearman rank correlation was 0.346 for more days in care.

- SMP members generally indicated that discrimination and calibration generally appear acceptable except for the highest days in care decile, raising issues related to outliers, (e.g., except for highest days in care decile). Members expressed concerns with the results from the excess days and mortality and the method of combining nursing home transitions.
- SMP members questioned whether there were meaningful differences in performance. The developer stated that a differences of three days should not be considered trivial from a cost perspective. But SMP members noted that it is not clear whether this equates to meaningful differences of quality of care, manifested for example, in differences in patient function or health-related quality of life. The same SMP member noted that the developer did not appear to test between vs. within ACO variance adjusted for risk factors. Another SMP member noted that a difference in three days could reflect variables not included in the risk adjustment model or in residual effects are not fully adjusted.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developers
- **Action items:**
  - The SMP should discuss the following:
    - Was the measure assessed and tested for an ACO and/or provider group level?
    - Is the risk adjustment approach sound and methodologically appropriate?
    - Is the exclusions list comprehensive?
    - Should the testing sample stretch beyond one year?

Subgroup 2

Measure# 0689: Percent of Residents Who Lose Too Much Weight (Long-Stay)

MEASURE HIGHLIGHTS

- Maintenance Measure
- **Description:** This measure reports the percentage of long-stay nursing home residents with a target Minimum Data Set (MDS) assessment (OBRA, PPS, Discharge) that indicates a weight loss of 5 percent or more of the baseline weight in the last 30 days or 10 percent or more of the baseline weight in the last six months, which is not a result of a physician-prescribed weight-loss regimen. The baseline weight is the resident's weight closest to 30 or 180 days before the date of the target assessment. Long-stay residents are identified as residents who have had at least 101 cumulative days of nursing facility care.
- **Type of measure:** Outcome
- **Data source:** Minimum Data Set MDS 3.0, collection instrument is the Resident Assessment Instrument (RAI)
- **Level of analysis:** Facility
- **Not risk-adjusted**
- **Sampling allowed**: N/A

- **Ratings for reliability:** 3 high 5 moderate 3 low and 0 insufficient → Measure passes with MODERATE rating
- Reviewers had no concerns with patient/encounter level reliability.
    - Critical data element reliability: Saliba et al. (2008) study looked at inter-rater reliability of the MDS elements (gold standard vs. gold standard; and gold standard vs local nurse).
    - Kappa analysis of gold standard nurse to facility nurse:
        - Weight loss: 0.92
        - Prognosis: 0.96
- Reliability testing conducted at the Accountable Entity level (facility):
    - Performance score reliability: The developer conducted both split-half reliability and usual beta-binomial signal-to-noise reliability testing.
        - Split-half reliability testing correlation was positive, and the relationship was moderate: ICC = 0.64; SNR = 0.76.
- There were some concerns with measure score reliability, though reviewers overall deemed it sufficient.
    - Measure score reliability testing was viewed as less compelling. Reviewers expressed a preference for the developer to have provided the distribution of the signal-to-noise reliability scores across facilities.
    - Several reviewers noted that less than one half of the facilities (44.6 percent) had no quarter-to-quarter change, while 20 percent had a >3 decile change quarter-to-quarter. Similarly, more than one third (37.4 percent) had a mean score where the 95 percent CI did not include the national mean.
- **Ratings for validity:** 1 high 5 moderate 5 low and 0 insufficient → Consensus not reached
- Reviewers had no concerns with critical data element validity.
    - Critical Data Element Testing: Relied on previous studies that have looked at inter-rater agreement. Examined a national validation of MDS 3.0 that tested the criterion validity of the items by examining the agreement between gold-standard nurse assessments and facility nurse assessments based on Kappa statistics. Kappa statistics ranged between 0.92-0.96, which are considered to be very good agreement.
- Validity testing conducted at the Accountable Entity level: facility
    - Performance Measure Score was tested using 1) Correlation with other measures of nursing facility quality including facility CMS five-star rating, health inspections rating, and staffing levels (overall and for RNs); and 2) seasonal variation.
        - Convergent validity testing was conducted. The correlation results show negative correlations between the facility-level weight loss QM score and the overall quality rating (ρ = -

.091, p < .0001), health inspection rating (ρ = -.056, p < 0.001), overall staffing level (ρ = -.041, p <0.0001), and RN staffing (ρ = -0.031, p=0.0001).
    - Seasonal variation showed highest weight loss in Q1 with progressively lower rates in Q2-Q4.
- Reviewers voiced concerns with convergent validity correlation results, citing weak negative correlations between the facility-level weight loss QM score and the overall quality rating.
- One reviewer noted that although low correlations are common, these are lower than what are typically seen, indicating that overall nursing home quality and staffing have little impact on residents' likelihood of losing weight. This may indicate the weight loss is more due to patient conditions that a nursing home has less control over and not the quality of care provided.
- Reviewers were also concerned with the developer's decision not to risk adjust.
- Risk adjustment was explored but the measure was not risk adjusted. The stated reason was that the developer's attempts to develop a risk adjusted model were unsuccessful, resulting in a low R-Squared value.
- The developer stated that risk adjustment was unsuccessful, but reviewers noted this may reflect the tight range of scores on this measure, which leads to questions as to relevance of this measure. If there are not specific risk factors that may lead to weight loss and could be addressed through appropriate interventions, is this a good quality measure? Literature indicates there are potentially addressable risk factors for unintentional weight loss in long term care facility residents, such as depression, cancer, Parkinson's disease, cognitive impairment, cardiac disorders, benign gastro diseases, eating dependencies, leaving 25 percent or more of food uneaten, and swallowing/chewing problems (all MDS scored items). Reviewers would have liked to see which covariates were tested in the risk adjustment model that had no predictive power at all as it is surprising that none of these factors were associated with weight loss.
- Comorbidities were not included in risk adjustment models. Age was specified as a linear variable. The association between weight loss and age is likely to be non-linear, and this should be explored. Patients with certain comorbidities (e.g., cancer) may be more likely to experience weight loss.

ITEMS TO BE DISCUSSED

- Additional clarifying information from the developers
- **Action items:**
    - The SMP needs to discuss and revote on the validity concerns, including the correlation testing results and the lack of risk adjustment.

- Does the SMP have concerns about the measure's ability to determine meaningful differences in performance, especially as related to variance, stability, and testing results?

Appendix A: Measures That Passed (Not Pulled for Discussion) (Detailed)

Subgroup 2

Measure# 3633e: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computer Tomography (CT) in Adults (Clinician Level)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This eCQM provides a standardized method for monitoring the performance of diagnostic CT to discourage unnecessarily high radiation doses, a risk factor for cancer, while preserving image quality. It is expressed as a percentage of eligible CT exams that are out-of-range based on having either excessive radiation dose or inadequate image quality, relative to evidence-based thresholds based on the clinical indication for the exam. All diagnostic CT exams of specified anatomic sites performed in inpatient, outpatient and ambulatory care settings are eligible.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Electronic Health Data; Electronic Health Records
- **Level of analysis:** Clinician: Individual
- **Risk-adjusted:** For each CT scan, a size-adjusted radiation dose is calculated based on the following items: 1) the actual radiation dose of the exam (unadjusted), the diameter of the anatomic area examined, the expected diameter based on the CT category, and a size adjustment coefficient of the CT category associated with the exam. This yield as a size-adjusted radiation dose for each CT. If either the size-adjusted radiation dose or the global noise (which is not size-adjusted) is out of range, the CT fails the measure.
- **Sampling allowed**: No
    - **Ratings for reliability:** 9 high 2 moderate 0 low and 0 insufficient → Measure passes with HIGH rating
        - Reliability testing conducted at the Accountable Entity level: clinician
        - Measure score reliability was estimated at the clinician level using the ICC, using randomly split samples for each accountable entity with 1,000 repetitions, applying a one-way random effects model, assuming that both entity effects and residual effects are random, independent, and normally distributed with mean 0. The Spearman-Brown prophecy formula was applied to adjust reliability from one-month test samples to the anticipated 12-month sample (i.e., (12*r)/(1 + (11*r)). These ICC (1) estimates (bounded between 0 and 1) were then logit-transformed and used to model the linear relationship between entity volume and logit reliability. By ranking predicted reliabilities across the complete range of potential volumes, the volume threshold that would correspond to ICC(1)=0.9 for an accountable entity was estimated.
        - The estimated mean split-half ICC was 0.99, using 47,635 CT exams collected from 606 individual clinicians (after Spearman-Brown adjustment to a 12-month data collection period). The number of exams per clinician in the one month of data used for testing ranged from 1 (which were excluded) to 604 (mean=77); predicted reliability for 12 months exceeded 0.90 for 89 percent of participating clinicians; 8 percent of individual clinicians in field-testing would not meet the minimum denominator to achieve ICC > 0.90.

- One reviewer stated that an ICC estimate greater than 0.90 may be interpreted as excellent reliability.
- Overall, reviewers viewed reliability testing as appropriate and results as acceptable.

- **Ratings for validity:** 5 high 6 moderate 0 low and 0 insufficient → Measure passes with MODERATE rating

- Validity testing conducted at the Accountable Entity level: clinician
    - Empirical Validity Testing: The results of the medical record review were compared with the results of the eCQM computation by selecting a sample of exams (N=8,000) representative of exams generated by the 606 individual clinicians across the eight health systems/vertically integrated organizations. The out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches.
    - Developer conducted data element validity:
    - Developers used criterion validity to compare agreement between the CT category (assigned using an algorithm assigning each CT exam to one of 18 CT categories based on ICD-10 and CPT codes) versus a gold standard method based on expert review of the complete medical record.
        - Results (weighted by the distribution of CT categories in the UCSF International CT Dose Registry): sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams). When tested across the 606 individual clinicians, the correct classification rate of the assignment of CT exams to CT category in field-testing was 95 percent on average. About 90 percent of tested individual clinicians had a correct classification rate of 80 percent or above. Most of the individual clinicians with correct classification rates below 80 percent had low sample sizes from the one month testing period (i.e., 5.1 percent read only one CT scan).
    - Patient size: Methods for measuring patient diameter on CT images have been previously validated including measuring patient size on axial and coronal images. Developer relied on published work and tested how often this method generated clinically plausible and non-missing values for size in testing data.
    - Radiation dose: The measure uses a standardized data element, generated by virtually (>99 percent) all CT machines, that is well validated and used broadly to reflect the radiation dose delivered to the patient. The proposed measure adjusted dose length product (DLP) for patient size to ensure that differences in patient mix would not result in differences in measure scores across reporting entities. Developers relied on this published work and tested how often this method generated clinically plausible and non-missing values for radiation dose in testing data.

- Size-adjusted radiation dose: When out-of-range rates are unadjusted for patient size, observed failure rates are strongly associated with size, with almost all failures occurring in larger patients. When failure rates are adjusted for size, there is no association. Using field testing data, developers assessed whether size-adjusted radiation dose could be calculated within a plausible range and quantified missing data.
    - Size-adjusted radiation dose: In field testing data this was within plausible range for 99 percent of CT exams and was missing for 0.4 percent of exams.
- Global noise: Adapted previously validated approaches. The developer assessed whether they could calculate global noise within a plausible range and quantified missing data using field-testing data.
    - Global noise was within a plausible range for 100 percent of CT exams in field-testing. Global noise was missing for 0.01 percent of examinations. The correlation between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams). There were four CT categories with exams in which global noise exceeded the allowable threshold. For other CT categories, exams were not observed above the threshold.
- Developer also calculated the correlation between global noise and physician dissatisfaction with image quality using data from the Image Quality Study and explored the rate of physician dissatisfaction in CT exams that exceeded global noise thresholds. Thresholds for "out-of-range" values to define numerator: Radiologists' satisfaction with CT images was used as a basis for establishing the maximum radiation dose and minimum image quality thresholds for each CT category.
- Measure score validity: tested using systematic assessment of face validity of measure score as an indicator of quality through a six-question poll to a TEP. The TEP represented a diverse group of clinicians (N=10), patient advocates (N=2), and leaders of medical specialty societies, payers, and healthcare safety and accrediting organizations. Face validity results were very strong with items having 100 percent agreement.

- Reviewer states that in spite of above reported results, at the individual clinician level, only 52 percent of participating clinicians would meet the threshold to detect an "out-of-range" prevalence five percentage points above the mean (i.e., 38 percent). Only 54 percent of participating clinicians would meet the threshold to detect an "out-of-range" prevalence five percentage points below the mean (i.e., 28 percent). To resolve this problem, the developers propose: (1) we measure users accept the ability to detect only larger deviations in performance; and (2) to set a minimum volume

threshold for reporting purposes. For example, a minimum annual volume of 145 CT scans (for reporting purposes) would provide 80% power to detect an "out-of-range" threshold either 10 percentage points above or below the mean (i.e., 23 percent or 43 percent) while excluding only 22 percent of participating clinicians, based on our test data.
    - Reviewer states that these limitations need to be clearly stated in the implementation specifications.
- Overall reviewers found validity testing to be appropriate and results to be acceptable.

Measure# 3662e: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Group Level)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This electronic eCQM provides a standardized method for monitoring the performance of diagnostic CT to discourage unnecessarily high radiation doses, a risk factor for cancer, while preserving image quality. It is expressed as a percentage of eligible CT exams that are out-of-range based on having either excessive radiation dose or inadequate image quality, relative to evidence-based thresholds based on the clinical indication for the exam. All diagnostic CT exams of specified anatomic sites performed in inpatient, outpatient, and ambulatory care settings are eligible.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Electronic Health Data; Electronic Health Records
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** For each CT scan, a size-adjusted radiation dose is calculated based on the following items: 1) the actual radiation dose of the exam (unadjusted), the diameter of the anatomic area examined, the expected diameter based on the CT category, and a size adjustment coefficient of the CT category associated with the exam. This yield as a size-adjusted radiation dose for each CT. If either the size-adjusted radiation dose or the global noise (which is not size-adjusted) is out of range, the CT fails the measure.
- **Sampling allowed**: No
    - **Ratings for reliability:** 8 high 3 moderate 0 low and 0 insufficient → Measure passes with HIGH rating
    - Reliability testing conducted at the Accountable Entity level:
        - Testing was performed at the clinician/group level in 16 groups within seven health systems and one vertically integrated organization.
        - Testing was conducted over four weeks.
        - The estimated mean split-half ICC using 48,500 CT exams was 0.99 (after Spearman-Brown adjustment to a 12-month data collection period).
        - The clinician groups ranged in size from 31 to 109 physicians (mean=27). The number of exams per clinician group in the one month of data used for testing ranged from 56 to 14,312 (mean=3,031).

    - Predicted reliability for 12 months exceeded 0.99 for every clinician group.
    - Based on this method, a minimum of 28 CT exams are required to achieve 90 percent reliability.
- Reviewers had some concerns about specifications:
    - One group had a high number of missing radiation doses (1,761) compared to non-missing radiation doses (6,157).
    - The measure is heavily dependent on proprietary software from the developer (UCSF and Alara imaging).
    - Time period for data collection was inconsistent.
- Reviewers had no concerns about the reliability testing.
- It was noted the reliability results demonstrated that reliability was very high.
- **Ratings for validity:** 7 high 4 moderate 0 low and 0 insufficient → Measure passes with HIGH rating
- Validity testing conducted at the Patient or Encounter level:
    - CT category - An ICD-10 based algorithm to assign the CT category was compared to chart review as the gold standard. The results, weighted by the distribution of CT categories in the UCSF International CT Dose Registry, were a sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams). When tested across the 16 clinician groups, the correct classification rate of the assignment of CT exams to CT category in field-testing was 92% on average and varied from 88-97 percent across the 16 clinician groups.
    - Patient size - A previously validated algorithm that used cross-sectional imaging to generate patient size estimates was compared to how often this method generated clinically plausible and non-missing data. Size-adjusted radiation dose could be calculated and was within plausible range for 99 percent of CT exams and was missing for 0.4 percent of exams.
    - Radiation dose - Dose-length product, which is generated by the CT machine for each examination, which relied on published work.
    - Size-adjusted radiated dose - When out-of-range rates are unadjusted for patient size, there are failure rates that are strongly associated with size, with almost all failures occurring in larger patients. When failure rates are adjusted for size, there is no association. Using field testing data, the developer assessed whether we could calculate size-adjusted radiation dose within a plausible range and quantified missing data. Size-adjusted radiation dose could be calculated and was within plausible range for 99 percent of CT exams and was missing for 0.4 percent of exams.
    - Global noise - The developer tested whether global noise could be calculated within a plausible range and quantified missing data. Global noise was also correlated with physician dissatisfaction with image quality. Global noise could be calculated and was within a plausible range for 100 percent of CT exams in field-testing. Global noise was missing for 0.01 percent of examinations. The correlation

between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams).
        - Thresholds for "out-of-range" values to define numerator - The developer used physician satisfaction with CT images as a basis for establishing the maximum radiation dose and minimum image quality thresholds for each CT category.
- Validity testing conducted at the Accountable Entity level:
    - eCQM output (encounter-level validity) was compared against medical record review using field testing data collected from eight health systems/vertically integrated organizations. The "medical record review" was a human-reviewed indicator of whether the size-adjusted radiation dose or global noise of each sampled exam exceeds predetermined thresholds, thus constituting a "gold standard." In a sample of 8,000 exams (1,000 per site), the out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches, indicating a correct and robust implementation of the measure logic.
    - Score-level testing, face validity as an indicator of quality
        - A six-question poll was posed to a TEP which represented a diverse group of clinicians (N=10), patient advocates (N=2), and leaders of medical specialty societies, payers, and healthcare safety and accrediting organizations. TEP members were identified by reaching out to key stakeholder organizations and advocates and identifying researchers who had contributed to the relevant literature.
        - 1. Do you agree that radiation dose is a relevant metric of quality for CT imaging? 100 percent agreement.
        - 2. Do you agree that image noise is a relevant metric of quality for CT imaging? 100 percent agreement.
        - 3. Do you agree that size is an appropriate method for adjusting for radiation dose for a given indication? 100 percent agreement.
        - 4. Do you agree that performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, is a representation of quality? 100 percent agreement.
        - 5. Do you agree that if this measure is implemented in the CMS hospital programs that this measure is likely to lead to reductions in radiation dose while maintaining adequate image quality? 100 percent agreement.
        - 6. How likely is it that implementation of this size-adjusted and stratified measure, as specified by the UC development team, in the Merit-based Incentive Payment System (MIPS), will lead to a reduction in average CT radiation dose while maintaining adequate CT image quality?

        - 16/17 members (94 percent) voted in favor: Five voted "very likely," and 11 voted "somewhat likely."
    - How likely is it that implementation of this size-adjusted and stratified measure, as specified by the UC development team, in the MIPS and hospital quality reporting programs (inpatient/outpatient), will lead to a reduction in average CT radiation dose while maintaining adequate CT image quality?
        - 16/17 members (94 percent) voted in favor: 10 voted "very likely," and six voted "somewhat likely."
- Reviewers noted that the face validity was very high.
- There were no concerns about the validity results except for the missing data comment (above).
- It was noted that a justification was not included to not adjust for social risk factors.

Measure# 3663e: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Level)

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This eCQM provides a standardized method for monitoring the performance of diagnostic CT to discourage unnecessarily high radiation doses, a risk factor for cancer, while preserving image quality. It is expressed as a percentage of eligible CT exams that are out-of-range based on having either excessive radiation dose or inadequate image quality, relative to evidence-based thresholds based on the clinical indication for the exam. All diagnostic CT exams of specified anatomic sites performed in inpatient and hospital outpatient care settings are eligible.
- **Type of measure:** Outcome: Intermediate Clinical Outcome
- **Data source:** Electronic Health Data; Electronic Health Records
- **Level of analysis:** Clinician: Individual
- **Risk-adjusted:** For each CT scan, a size-adjusted radiation dose is calculated based on the following items: 1) the actual radiation dose of the exam (unadjusted), the diameter of the anatomic area examined, the expected diameter based on the CT category, and a size adjustment coefficient of the CT category associated with the exam. This yield as a size-adjusted radiation dose for each CT. If either the size-adjusted radiation dose or the global noise (which is not size-adjusted) is out of range, the CT fails the measure.
- **Sampling allowed**: No
    - **Ratings for reliability:** 9 high 2 moderate 0 low and 0 insufficient → Measure passes with HIGH rating
    - Reliability testing conducted at the Accountable Entity level:
        - Reliability testing was conducted in electronic health records from 2/20 to 4/21.
        - Testing performed at the facility level in 16 hospitals within seven health systems and one vertically integrated organization.
        - The estimated mean split-half ICC using 37,172 CT exams was 0.99. The number of exams per hospital in the one month of data used

for testing ranged from 625 to 6,157 (mean=2,323); predicted reliability for 12 months exceeded 0.99 for every hospital.
    - The number of CT exams obtained during inpatient hospitalizations (n=15) in the one month of testing data ranged from 134-1,568 (mean 715); thus, the number of CT exams from inpatient settings per hospital is estimated to vary from 1,608-18,816 for a 12-month period. For the individual hospitals, the predicted reliability for 12 months of inpatient CT exams exceeded 0.99 for every hospital during the testing phase.
    - The number of CT exams obtained during hospital outpatient encounters, including ED encounters, in the one month of testing data ranged from 119-4,978 (mean 1,608); thus, the number of CT exams from outpatient settings per hospital is estimated to vary from 1,428-59,736 for a 12-month period. For the individual hospitals, the predicted reliability for 12 months of outpatient CT exams exceeded 0.99 for every hospital during the testing phase.
- Reviewers had some concerns about specifications:
    - One group had a high number of missing radiation doses (1,761) compared to non-missing radiation doses (6,157).
    - The measure is heavily dependent on proprietary software from the developer (UCSF and Alara imaging).
    - Time period for data collection was inconsistent.
- Reviewers had no concerns about the reliability testing.
- It was noted the reliability results demonstrated that reliability was very high.
- **Ratings for validity:** 6 high 5 moderate 0 low and 0 insufficient → Measure passes with HIGH rating
- Validity testing conducted at the Patient or Encounter level:
    - CT category - An ICD-10 based algorithm to assign the CT category was compared to chart review as the gold standard. The results, weighted by the distribution of CT categories in the UCSF International CT Dose Registry, were a sensitivity = 0.86 and specificity = 0.96 (n=978 CT exams). When tested across the 16 hospitals, the correct classification rate of the assignment of CT exams to CT category in field-testing was 92 percent on average and varied from 88-97 percent.
    - Size-adjusted radiated dose - When out-of-range rates are unadjusted for patient size, there are failure rates that are strongly associated with size, with almost all failures occurring in larger patients. When failure rates are adjusted for size, there is no association. Using field testing data, the developer assessed whether we could calculate size-adjusted radiation dose within a plausible range and quantified missing data. Size-adjusted radiation dose could be calculated and was within plausible range for 99 percent of CT exams and was missing for 0.4 percent of exams.
    - Global noise - The developer tested whether global noise could be calculated within a plausible range and quantified missing data.

Global noise was also correlated with physician dissatisfaction with image quality. Global noise could be calculated and was within a plausible range for 100% of CT exams in field-testing. Global noise was missing for 0.01% of examinations. The correlation between noise and physician dissatisfaction with image quality is 0.37 overall based on the image quality study (n=727 CT exams).

- Validity testing conducted at the Accountable Entity level:
  - Score-level testing, empirical validity testing: Gold standard comparison:
    - eCQM output (encounter-level validity) was compared against medical record review using field testing data collected from eight health systems/vertically integrated organizations. The "medical record review" was a human-reviewed indicator of whether the size-adjusted radiation dose or global noise of each sampled exam exceeds predetermined thresholds, thus constituting a "gold standard." In a sample of 7,000 exams (1,000 per site), the out-of-range results (measure score) from the medical record review and the eCQM computation were identical with no discrepancies between the two approaches, indicating a correct and robust implementation of the measure logic.
  - Score-level testing: Face validity
    - A poll was posed to a TEP representing a diverse group of clinicians (N=10), patient advocates (N=2), and leaders of medical specialty societies, payers, and healthcare safety and accrediting organizations.
    - 1. Do you agree that radiation dose is a relevant metric of quality for CT imaging? 100 percent agreement.
    - 2. Do you agree that image noise is a relevant metric of quality for CT imaging? 100 percent agreement.
    - 3. Do you agree that size is an appropriate method for adjusting for radiation dose for a given indication? 100 percent agreement.
    - 4. Do you agree that performance on this measure of radiation dose and image quality, adjusted for size, stratified by indication, is a representation of quality? 100 percent agreement
    - 5. Do you agree that if this measure is implemented in the CMS hospital programs that this measure is likely to lead to reductions in radiation dose while maintaining adequate image quality? 100 percent agreement.
    - 6. How likely is it that implementation of this size-adjusted and stratified measure, as specified by the UC development team, in the Merit-based Incentive Payment System (MIPS), will lead to a reduction in average CT radiation dose while maintaining adequate CT image quality?

        - 16/17 members (94 percent) voted in favor: Five voted "very likely," and 11 voted "somewhat likely."
    - How likely is it that implementation of this size-adjusted and stratified measure, as specified by the UC development team, in the MIPS and hospital quality reporting programs (inpatient/outpatient), will lead to a reduction in average CT radiation dose while maintaining adequate CT image quality?
        - 16/17 members (94 percent) voted in favor: 10 voted "very likely," and six voted "somewhat likely.
- Reviewers noted that the face validity was very high.
- There were no concerns about the validity results except for the missing data comment (above).
- It was noted that a justification was not included to not adjust for social risk factors.

Measure# 3665: Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood

MEASURE HIGHLIGHTS

- New Measure
- **Description:** This is a multi-item measure consisting of four items: Q1: "I felt heard and understood by this provider and team," Q2: "I felt this provider and team put my best interests first when making recommendations about my care," Q3: "I felt this provider and team saw me as a person, not just someone with a medical problem," Q4: "I felt this provider and team understood what is important to me in my life."
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Electronic Health Records; Instrument-Based Data
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with two risk factors (i.e., survey mode and proxy-assist)
- **Sampling allowed**: The target population for sampling includes patients aged 18 years or older who received ambulatory palliative care services from a MIPS-eligible provider within the three months prior to the start of survey fielding. Findings from the alpha pilot test and beta field test support the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. The provider or program will provide a vendor with an extract file of all patients who received care during the measurement period. To prevent gaming and to minimize administration and social desirability bias, the vendor will apply the eligibility criteria to identify the patient sample and field the survey to eligible patients. Survey administration will be mixed-mode, including web (emailed link to online survey), mail (hard-copy of the survey) followed by telephone (Computer Assisted Telephone Interviewing) survey if needed.
  Assessments of measure reliability based on the ICC suggest that programs will need a sufficient sample to have at least 37 completed responses to the *Feeling Heard and Understood* items over the 12-month reporting period.
    - **Ratings for reliability:** 3 high 6 moderate 1 low and 1 insufficient → Measure passes with MODERATE rating
    - Reliability Specifications:

- Consider adding “who completed the survey” to the end denominator statement.
- The description does not include the target population, timeframe, age range, and setting.
- Should telehealth be considered an eligible encounter for the denominator?
- The MIPS provider is identified based on a three-month period versus the survey that refers to "the last six months". Further, the highest volume three-month and six-month providers may not be the same. The survey does not identify eligible ambulatory palliative care, and patients who transfer to hospice may also complete the survey with hospice treatment responses. The patient’s ability to differentiate ambulatory and non-ambulatory palliative care is not well defined. These items may lead to potential misattribution issues.
- The submission states that surveys completely filled out by a proxy are excluded. However, it is unclear how this would be identified. Question 11 of the survey states "answered the questions for me" which is assumed to signify that the patient was not involved and is not well defined. It is recommended the language be clarified before use.
- Clarify if the measure reports the top-box results for each of the four survey questions, and the average top-box scores as both would be used for targeted improvement activities. Clear calculation details are not apparent, specifically the conversion for the 1-100 rate.

- Reliability testing conducted at the Patient or Encounter level:
  - The multi-data four-question scale was evaluated with Cronbach’s alpha with an acceptable threshold of 0.7. The four-data elements of the *Feeling Heard and Understood* scale Cronbach’s were 0.90.
  - A test-retest reliability coefficients calculation of live phone respondents in high-volume ambulatory programs were given a shortened computer-assisted telephone interview (CATI) survey within 48 hours of the first survey. Only A reliability of at least 0.70 demonstrates acceptable reliability. The result from the polychoric correlation coefficient was 0.85 for the CATI data collection method. Agreement statistics were not provided.
  - It is not clear if the top-box approach was employed in patient or encounter level reliability testing.
  - Not all mixed-mode administration methods were used in the test-retest design. NQF’s measure evaluation guidance states, “If multiple data sources (i.e., instruments, methods, modes, languages) are used, then comparability or equivalency of performance measure scores should be demonstrated.”
  - “For item #2 is double barreled and research indicates that this leads to measurement error.”
- Reliability testing conducted at the Accountable Entity level:

- Using a signal-to-noise analysis, accountable entity testing was conducted to assess between- (i.e., signal) and within- (i.e., noise) subject variability to discriminate provider performance.
- Developers used hierarchical generalized-linear regressions to decompose variability of binomial outcomes to programs, and to covariates with the data hierarchy as patient observations. The variance of the model can be decomposed using the (adjusted) ICC, which provides a summary of the reliability of the measure as tested, with higher values implying more variability between programs. Using Bayesian generalized mixed-effects models obtained a posterior distribution of the adjusted ICC with estimates of 0.052 (95% CI: 0.027 to 0.089). The SMP members acknowledge that testing during the COVID-19 pandemic may have affected changes in palliative care services and experiences.
- For projected to observed variance from within each program, Spearman-Brown prophecy formula was used to determine reliability results to future samples. To obtain a result of 0.7 or higher, an average of 45 eligible and complete returned responses were required. Assuming high correlation between the four survey questions, as 3.25 estimated design effect from repeated measure, an average sample size of 37 eligible and complete respondents would be required. The 3.25 estimated design effect method description is not clear.
- To assess the average adjusted reliability of individual programs, developers estimated a posterior distribution for the overall variability using an Adams-like (2009) approach, which demonstrated an average reliability across programs of approximately r = 0.752. Is this figure "falsely elevated due to the absence of meaningful risk adjustment?"

- **Ratings for validity:** 3 high 5 moderate 3 low and 0 insufficient → Measure passes with MODERATE rating
- Validity testing conducted at the Patient or Encounter level:
    - Convergent validity testing was used for patient- encounter-level validity testing hypothesizing the relationship so similar constructs, including data elements from other instruments: 1) Consumer Assessment of Healthcare Providers and Systems [CAHPS] Hospice, (i.e., "In the last 3 months, how often did this provider and team listen carefully to you?") from the four-item CAHPS Communication composite measure, and 2) *3666 Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain,* "In the last 6 months, did you get as much help as you wanted for your pain from this provider and team?" This 3666-scale item is from a new NQF measure developed by the same developer that is currently being reviewed by SMP. Developers hypothesized feeling heard and understood would correlate to getting help for pain needs from the palliative care team. Interpretation of the bivariate correlation

followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50) (Rosnow & Rosenthal, 1989).
    - The *Feeling Heard and Understood* scale was associated with higher CAHPS communication scores (r = 0.54, p<.001). For *3666 Receiving Desired Help for Pain,* the correlations were weak/low (r = 0.48, p< .001). positive convergent validity for correlation coefficients is generally above .50, although usually recommended at higher rates, such as above 0.70, depending on intended use.
- Validity testing conducted at the Accountable Entity level:
    - To assess accountable entity level validity, measure scores examined the association of the measure scores to 1) the current NQF submitted *3666 Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain,* 2) the CAHPS communication measure score, and 3) the individual's overall rating of their palliative care provider and team. The source and collection method for the third correlate is clear. The developer hypothesized these scores would be positively associated to 3665. The measure was positively associated with the CAHPS communication quality measure (r = 0.635, p =0.011), the *Receiving Desired Help for Pain* quality measure (r = 0.496, p<.001) and the overall rating of the palliative care provider and team (r= 0.768, p=<.001) with associated to other similar measures (r = 0.5 – 0.8). Positive correlations between 3665 and 3666 *Receiving Desired Help for Pain* were moderate low.
    - Face validity was assessed with a panel of seven palliative care communication experts who assessed the final measure specifications and testing results and rate the measure's ability to distinguish quality palliative care. Face validity ratings were from 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9). The average face validity ratings of the measure score were 8.3 which corresponds to a developer defined average rating of "high."
    - Meaningful differences in performance were tested statistical differences using an ANOVA-like analysis using "full" (i.e., assuming at least one difference among programs) to "reduced/nested" (i.e., assuming no differences among programs) models. Using a ranking approach of all tested programs, equivalence of a difference in measure score to ranking of program performance were estimated and evidenced demonstrating diversity in program measure scores. "Full" models also demonstrated distinguishability among programs that were significantly different from the "nested" models without one ($c^2_{(43)}$=60.04, *p*<.05). Magnitude of change, rank order, distributions, and medians of performance are also provided, and demonstrate actionability of the measure results.
    - The developer states missing data is accommodated among the four items of the scale for adjusted score estimates and require no outcome imputation. Survey response rates, frequency, and analysis

were provided by age, gender, and race. Of the 7,595 fielded surveys in 10 rounds, 3,356 we not returned, 1,435 were excluded due to sample size, and 2,804 were included in the analysis, which equates to a 37 percent raw response rate and a 46 percent response rate excluding ineligible patients. Non-response rates between programs and demographic differences should be further explored. Developers report the mean item-level of missing data was 0.8% with low levels of missingness and no discernible patterns and no evidence of systematic bias. The sampled respondents versus non-respondents were heavily female (56.2 percent vs 54.5 percent; p=0.21), older (mean age 63.4 vs 60.9; p<0.01), and White (88.1 percent vs 80.2 percent; p<0.01).

- The submission details five specification exclusions and outlines an analysis that only examined the fifth listed exclusion, "Patients for whom a proxy completed the entire survey on their behalf for any reason (no patient involvement", n=435). The mean and standard deviation for the proxy-assist [0.77 (0.34)] was statistically different from the patient-only [0.71 (0.37)] and proxy-only [0.69 (0.37)] respondents. The one-way ANOVA among the three means was statistically different. The developers excluded the proxy-only.
- The measure is risk adjusted for survey mode and proxy assist using a hierarchical generalized-linear model that relates the proportion of top-box patient-level outcome responses to provider scores (conditioned on risk adjustment covariates) of patient observations in tested programs. Data for the hierarchical approach is assessed for a 12-month, while the measure captures survey responses for a three-month period.
- Patient, survey, and proxy factors were considered, including survey data or program information, provider data, and census data. Other examined and excluded risk adjustment factors included patient age; education; financial, physical (including primary hospice diagnosis), and mental health; and race and ethnicity.
- Risk adjustment statistical associations to the measure and programs were set at the 0.5 level of significance. Fisher's exact test assessed binary or categorical variables to the measure and variable differences across programs. The Z-test for a generalized linear model tested the association of the measure, and ANOVA F-test assessed differences between programs. P-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons to control the false discovery rate (Benjamini & Hochberg, 1995). The survey mode was the only variable to demonstrate statistically significant associations with both the measure (p=0.013) and with programs (p=0.001). Including proxy-assist in the risk model was determined by the Kendall's τ Test Statistic (0.88) and TEP guidance.

Measure #3666: Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain

MEASURE HIGHLIGHTS

- New Measure
- **Description:** The percentage of patients aged 18 years and older who had an ambulatory palliative care visit and report getting the help they wanted for their pain from their palliative care provider and team within six months of the ambulatory palliative care visit.
- **Type of measure:** Outcome: PRO-PM
- **Data source:** Instrument-Based Data
- **Level of analysis:** Clinician: Group/Practice
- **Risk-adjusted:** Statistical risk model with two risk factors (i.e., survey mode and proxy-assist)
- **Sampling allowed**: The target population for sampling includes patients aged 18 years or older who received ambulatory palliative care services from a MIPS-eligible provider within the three months prior to the start of survey fielding. Findings from the alpha pilot test and beta field test support the feasibility of identifying eligible patients using administrative data and using a survey vendor to support survey administration and data collection. The provider or program will provide a vendor with an extract file of all patients who received care during the measurement period. To prevent gaming and to minimize administration and social desirability bias, the vendor will apply the eligibility criteria to identify the patient sample and field the survey to eligible patients. Survey administration will be mixed-mode, including web (emailed link to online survey), mail (hard-copy of the survey) followed by telephone (Computer Assisted Telephone Interviewing) survey if needed.
  Assessments of measure reliability based on the ICC suggest that programs will need a sufficient sample to have at least approximately 33 completed responses to the *Receiving Desired Help for Pain* items over the 12-month reporting period.
    - **Ratings for reliability:** 4 high 5 moderate 2 low and 0 insufficient → Measure passes with MODERATE rating
    - Reliability Specifications
        - The MIPS provider is identified based on a three-month period versus the survey that refers to "the last 6 months." Further, the highest volume three-month and six-month providers may not be the same. The survey does not identify eligible ambulatory palliative care, and patients who transfer to hospice may also complete the survey with hospice treatment responses. The patient's ability to differentiate ambulatory and non-ambulatory palliative care is not well defined. These items may lead to potential misattribution issues.
        - Consider adding "who completed the survey" to the end denominator statement.
        - The data sources should include an EHR.
        - Are all eligible providers listed in the measure responsible for pain management in the palliative care patient? Considering the dramatic changes in care delivery in the last two years, should telehealth be added for future iterations of the measure?
        - It is not clear how patients differentiate palliative care providers who are responsible for pain management in the fourth exclusion.

Should an exclusion be added for providers who are not responsible for pain management?
    - To determine an adequate recall and survey response time frames, did the assessment of the evidence include the impacts of potential delays in transitioning from palliative to hospice care on pain management and relief?
    - It is unclear how a rate measure gets converted to a 1-100 scale. An example calculation would be helpful for the reader.
    - To achieve this n, the measured entity will have a much higher number of patients who are receiving ambulatory palliative care (n = 106 – 132) and receive the questionnaire but are filtered out prior to key question, based on lack of presence of pain or desire to have pain treated.
- Reliability testing conducted at the Patient or Encounter level:
    - Developers used a test-retest reliability coefficient and percent agreement to test the survey data element *Receiving Desired Help for Pain*. Although a mixed response modes are used for data collection (i.e., web (emailed link to online survey), mail (hard-copy of the survey) followed by telephone (Computer Assisted Telephone Interviewing) survey), only phone survey test respondents were eligible for CATI retest. The result from the polychoric correlation coefficient was 0.90 with 88% agreement for the CATI data collection method.
    - Table 2 (pg. 46) shows that there were approximately 2,800 completed and usable surveys in their beta (field) test and approximately 4,800 unusable surveys due to lack of response or ineligibility. Is there the possibility of a response bias that is either positive or negative in the usable data? Demographic data in tables that follow suggest that those completing the survey were likely to be white (~88 percent) and college educated (~65 percent).
- Reliability testing conducted at the Accountable Entity level:
    - Using a signal-to-noise analysis, accountable entity testing was conducted to assess between- (i.e., signal) and within- (i.e., noise) subject variability to discriminate provider performance.
    - Developers used hierarchical generalized-linear regressions to decompose variability of binomial outcomes to programs and to covariates with the data hierarchy as patient observations. The variance of the model can be decomposed using the (adjusted) ICC, which provides a summary of the reliability of the measure as tested, with higher values implying more variability between programs. Using Bayesian generalized mixed-effects models obtained a posterior distribution of the adjusted ICC with estimates of approximately 0.079 (95 percent CI: 0.02, 0.175) is “extremely low and is concerning.”
    - For projected to observed variance from within each program, Spearman-Brown prophecy formula was used to determine reliability results to future samples. To obtain a result of 0.7 or

higher, an average of 49 eligible and complete responses were required. Results across all programs r=0.482 and across programs with a minimum of 33 respondents (considering 68 percent response rates) were r=0.735. The SMP acknowledges that testing during the COVID-19 pandemic may have affected changes in palliative care services and experiences.
    - To assess the average adjusted reliability of individual programs, developers estimated a posterior distribution for the overall variability using an Adams-like (2009) approach, which resulted in an average reliability across programs of approximately r = 0.752.
- **Ratings for validity:** 2 high 6 moderate 3 low and 0 insufficient → Measure passes with MODERATE rating
- Validity testing conducted at the Patient or Encounter level:
    - Convergent validity testing was used for patient- encounter-level validity testing hypothesizing the relationship so similar constructs, including data elements from other instruments: 1) Consumer Assessment of Healthcare Providers and Systems [CAHPS] Hospice, (i.e., "In the last 3 months, how often did this provider and team listen carefully to you?") from the four-item CAHPS Communication composite measure, and 2) 3666 *Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain*, "In the last 3 months, how often did this provider and team listen carefully to you?" The developers also assessed the first data element of the 3665 four-item scale from a new NQF measure developed by the same developer and currently being reviewed in this cycle. Developers hypothesized that pain management would link with feeling heard and understood by that same palliative care provider and team. Interpretation of the bivariate correlation followed standard conventions for small, medium, and large associations (i.e., 0.10, 0.30, 0.50) (Rosnow & Rosenthal, 1989). Both showed moderate correlations. Feeling Heard and Understood scale were associated with higher CAHPS communication scores (r = 0.57, p<.001) and 3665 Feeling Heard and Understood, the correlations were weak/low (r = 0.61, p< .001). positive convergent validity for correlation coefficients is generally above .50, although usually recommended.
- Validity testing conducted at the Accountable Entity level:
    - To assess accountable entity level validity, measure scores examined the association of the measure scores to 1) the current NQF-submitted #3665 *Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood,* 2) the CAHPS communication measure score, and 3) the individual's overall rating of their palliative care provider and team. The source and collection method for the third correlate is clear. The developer hypothesized these scores would be positively associated to 3665. The measure showed low/weak positively associated with the CAHPS communication quality measure (r = 0.386, p =0.014) and the

Experience of Feeling Heard and Understood quality measure (r = 0.41, p<.009). It also showed moderate low linkage to overall rating of the palliative care provider and team (r= 0.56, p=<.001) with associated to other similar measures (r = 0.5 – 0.8).

- Face validity was assessed with a panel of seven palliative care communication experts who assessed the final measure specifications and testing results and rate the measure's ability to distinguish quality palliative care. Face validity ratings were from 1 (lowest rating) to 9 (highest rating); numeric ratings corresponded with descriptive ratings of low (1-3), moderate (4-6), or high (7-9). The average face validity ratings of the measure score were 7.7 which corresponds to a developer defined average rating of "high."
- Meaningful differences in performance were tested statistical differences using an ANOVA-like analysis using "full" (i.e., assuming at least one difference among programs) to "reduced/nested" (i.e., assuming no differences among programs) models. Using a ranking approach of all tested programs, equivalence of a difference in measure score to ranking of program performance were estimated and evidenced demonstrating diversity in program measure scores. "Full" models also demonstrated distinguishability among programs that were significantly different from the "nested" models without one ($c^2_{(42)}$=98.99, *p*<.05). Magnitude of change, rank order, distributions, and medians of performance are also provided, and demonstrate actionability of the measure results.
- Survey response rates, frequency, and analysis were provided by age, gender, and race. The sampled respondents versus non-respondents were heavily female (56.2 percent vs 54.5 percent; p=0.21), older (mean age 63.4 vs 60.9; p<0.01), and White (88.1 percent vs 80.2 percent; p<0.01).
- Missing data was assessed by non-item response and non-survey response. Of the 7,595 fielded surveys in 10 rounds, 3,356 we not returned, 1,435 were excluded due to sample size, and 2,804 were included in the analysis, which equates to a 37 percent raw response rate and a 46 percent response rate excluding ineligible patients. Non-response rates between programs and demographic differences should be further explored. Developers report the mean item-level of missing data was 0.8 percent with low levels of missingness and no discernible patterns and no evidence of systematic bias.
- The submission details five specification exclusions and outlines an analysis that only examined the fifth listed exclusion, "Patients for whom a proxy completed the entire survey on their behalf for any reason (no patient involvement", n=255). The mean and standard deviation for the proxy-assist [0.83 (0.37)] was not statistically different from the patient-only [0.79 (0.41)] and proxy-only [0.69 (0.37)] respondents. The one-way ANOVA among the three means was not significant. The developers excluded the proxy-only.

- The measure is risk adjusted for survey mode and proxy assist using a hierarchical generalized-linear model that relates the proportion of top-box patient-level outcome responses to provider scores (conditioned on risk adjustment covariates) of patient observations in tested programs. Data for the hierarchical approach is assessed for a 12-month, while the measure captures survey responses for a three-month period.
- Patient, survey, and proxy factors were considered including survey data or program information, provider data, and census data. Other examined and excluded risk adjustment factors included patient age; education; financial, physical (including primary hospice diagnosis), and mental health; and race and ethnicity.
- Risk adjustment statistical associations to the measure and programs were set at the 0.5 level of significance. Fisher's exact test assessed binary or categorical variables to the measure and variable differences across programs. The Z-test for a generalized linear model tested the association of the measure, and ANOVA F-test assessed differences between programs. P-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons to control the false discovery rate (Benjamini & Hochberg, 1995). No variables demonstrated statistically significant associations with the measure or programs. Including proxy-assist in the risk model was determined by the Kendall's τ Test Statistic (0.79) and TEP guidance. Pain was significantly tied to diagnostic group, although the developers report data quality challenges hindered an assessment.

Appendix B: Additional Information Submitted by Developers for Consideration

Subgroup 1

Measure Number: 3649e

Measure Title: Risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) electronic clinical quality measure (eCQM)

Measure Developer/Steward: Brigham and Women's Hospital

Reliability

- **Issue 1:** For element-level reliability, the developers presented provider-level summaries of the patient demographics across test and validation samples. It wasn't clear to me how this sheds light on element-level reliability.
    - **Developer Response 1:** We tested measure score reliability using the NQF eCQM Feasibility Scorecard (where all data elements scored a 1/1 for availability, accuracy, data standards, and workflow), and by using a test-retest method to assess the predicted/expected ratios at the clinician group level. Additionally, we performed a Spearman rank correlation to estimate how the test and validation samples agree with each other (how reliable they are), an ICC to assess variability across clinician groups, a Hosmer-Lemeshow calibration to assess the goodness of fit for the logistic regression model, and a C-statistic to assess the predictive strength of the model. Based on NQF standards, we have performed sufficient testing to assess the reliability of the RSCR eCQM as a new measure.
- **Issue 2:** Results from 17 providers may not be generalizable / concern that measure was tested in two EHR systems (Epic, Cerner).
    - **Developer Response 2:** We tested this measure across 17 clinician groups in two geographically different healthcare systems with commonly used vendors, Epic and Cerner. Measure testing in additional sites is costly and burdensome, the sample size and homogeneity of patients is noted as a limitation of this measure. Future analysis with additional healthcare systems is expected to increase variation in risk adjusted RSCR rates. As this measure is retooled from the NQF1550 and NQF3493 with an expanded inclusion criteria (patients 18 years and older, inpatient and outpatient procedures and complications, all payer data), we expect to see even more variation in rates upon measure implementation compared to these existing measures.
- **Issue 3:** Measurement window when the measure is reported, impact on Spearman correlation.
    - **Developer Response 3:** Upon implementation, this eCQM is designed to report annual rates, although this analysis used data from across four years. If implemented, this measure would use a national sample and annual analysis is not expected to differ significantly. This measure is retooled from the NQF1550 and the NQF3493 CQMs, which are used to report the Hospital Compare THA and TKA RSCRs for Medicare beneficiaries receiving inpatient procedures. As Hospital Compare reports on a smaller population than our measure annually, we do not expect to encounter any difficulties in

reporting rates by year rather than across years. Additionally, we do not expect that this had an impact on our Spearman correlation.

- **Issue 4:** Interpretation of the ICC and version of the ICC used.
    - **Developer Response 4:** The ICC of 0.006 reflects the variation in patient score attributed to providers, a measure which tends to be <5%. This low ICC is likely attributed to the small sample size; inclusion of additional clinician groups in analysis is expected to raise the ICC and decrease the range of the confidence intervals.
    - An ICC was calculated to describe how much variation in the provider-level scores is due to provider-level signal variation in the 2019 sample, resulting in an ICC of 0.7261, well above the NQF required 0.50. We chose to only include 2019 in this ICC to demonstrate how the measure can meaningfully be used to report data annually. An ICC with all four years of data collected is expected to be higher.
    - As noted by Reviewer 11 (question 11), the strong split-half results outweigh the ICC results, which have room for improvement.
- **Issue 5:** Asking for Table 9 results
    - **Developer Response 5:** This comment refers to the following table:

Risk Adjusted Results and Confidence Intervals:

*	*	Test	Test	Validation	Validation
Site	**Clinician Group**	**Adjusted Rate**	**95% CI**	**Adjusted Rate**	**95% CI**
MGB	A	3.792	3.12-4.47	3.777	3.14-4.41
MGB	B	3.415	2.56-4.26	3.451	2.61-4.29
MGB	C	3.722	3.02-4.43	3.663	2.98-4.35
MGB	D	3.561	2.67-4.45	3.532	2.66-4.40
MGB	E	3.891	3.11-4.67	3.913	3.14-4.68
MGB	F	3.235	2.55-3.92	3.298	2.63-3.96
Cerner	A	N/A	N/A	N/A	N/A
Cerner	B	3.631	2.74-4.53	3.693	2.78-4.6
Cerner	C	4.220	3.35-5.09	4.279	3.42-5.14
Cerner	D	3.814	3.16-4.47	3.850	3.15-4.55
Cerner	E	3.675	2.94-4.41	3.700	2.97-4.42
Cerner	F	3.583	2.76-4.41	3.576	2.83-4.33
Cerner	G	3.766	3.06-4.47	3.733	3.08-4.38
Cerner	H	3.600	2.83-4.37	3.561	2.85-4.27
Cerner	I	3.506	2.75-4.26	3.536	2.81-4.26
Cerner	J	3.773	2.92-4.63	3.792	2.94-4.65
Cerner	K	3.547	2.77-4.32	3.514	2.80-4.22
Cerner	L	3.506	2.77-4.24	3.422	2.72-4.12

***This cell is intentionally left empty**

Validity

- **Issue 1:** Defining a gold standard group for comparison, inter-rater reliability of data elements.
    - **Developer Response 1:** The BWH team chose not to label either the manual chart review or the eCQM output as the "gold" standard as both methods demonstrated room for inaccuracies during the testing process. The BWH team has tested the data abstraction accuracy of the eCQM, and we have shown that it reliably pulls the correct information from the EHR. We acknowledge that the eCQM relies on the correct input

of information into the EHR. Conversely, manual chart review and the ability to read through progress notes inherently provides more qualitative information than what can be captured by an eCQM. However, with manual review of the EHR data, at times BWH reviewers would incorrectly interpret information because procedures/diagnostic codes are occasionally buried or hidden within the EHR. Due to the drawbacks of both methods, the BWH team decided that reporting a kappa value evaluating the similarity of results between both methods would be more appropriate than generating a PPV and/or NPV.

- Regarding inter rater reliability for data element, and comments that we presented data presence rather than accuracy, inter-rater reliability was performed using a random sample of 30 BWH patients and passed the 70% agreement threshold.

- **Issue 2:** Concern that the TEP only consisted of seven members, and a single voting question**.**
  - **Developer Response 2:** Our TEP was comprised of orthopedic surgeons, clinicians, patient representatives, and measure development professionals. Although the face validity vote itself was comprised of a single question (**"***The performance scores resulting from the risk-standardized complication rate eCQM, as specified, can be used to distinguish good from poor clinician group-level quality related to patient safety*.") it is backed by continuous meetings with the TEP over the course of 3 years throughout measure development where the TEP had an opportunity to discuss ongoing development of measure specifications and analysis, provide feedback, and ask questions.
- **Issue 3:** Concern over lack of variation in rates
  - **Developer Response 3:** Our rates ranged from 3.08%-3.95% and demonstrated similar variation to existing metrics which incorporate larger samples. This measure is a retooled version of the NQF1550 and NQF3493 RSCRs. The rates of this measure are shown in the testing paper published by Bozic et al., 2014, and are also reported on an ongoing basis through the CMS Hospital Compare measure. These measures only include Medicare beneficiaries aged 65 years and older with procedures and complications associated with the inpatient setting. These are also CQMs rather than eCQMs, meaning they used claims-based data rather than EHR data, which is associated with a delay in reporting. Bozic et al., 2014 showed a median risk standardized complication rate of 3.6% across over 878,098 patients with rates ranging from 1.8%-9.0%. Since 2014, this median complication rate has gone down to 2.4%, as shown in the Hospital Compare measure national observed rate. In comparison to the national rate, the rates across MGB and the Cerner site are described as "no different from the national rate" as they fall within the 95% interval estimates of the national rate. In comparison to the Hospital Compare rate, we expected the risk adjusted rates of the proposed RSCR eCQM to be higher than the current national observed rate considering that this measure includes outpatient procedures and complications, where the NQF1550 CQM is limited to procedures and complications associated with the inpatient setting.
    - Bozic KJ, Grosso LM, Lin Z, et al. (2014). "Variationinhospital-levelrisk-standardizedcomplicationratesfollowingelective primary total hip and knee arthroplasty." J Bone Joint Surg Am 96:640-7.

Other General Comments

Measure Specifications:

- **Issue 1:** Coding is submitted as SNOMED CT codes- unclear if those are generally available for measurement.
    - **Developer Response 1:** The codes sets for all of our measures are published in the Value Set Authority Center (VSAC) and use standard codes that are recommended by CMS Blueprint and Promoting Interoperability program (see table below displaying query from NLM's VSAC):

Table 1: Brigham and Women's Hospital Published Code Sets

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Opioid Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.46	*	20210723	07/23/2021	Published
Total Knee Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.33	*	20210311	03/11/2021	Published
Total Hip Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.32	*	20210311	03/11/2021	Published
Active Bleeding	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.28	*	20201122	11/22/2020	Published
Anticoagulant Medications, Oral	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.20	*	20200804	08/04/2020	Published
Anticoagulant Medications, Injection	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.21	*	20200804	08/04/2020	Published
Anticoagulant Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.19	*	20200714	07/14/2020	Published
Acute Respiratory Failure	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.18	*	20200708	07/09/2020	Published
Coagulation Disorder Conditions	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.17	*	20200428	04/28/2020	Published
General Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.15	*	20200424	04/24/2020	Published
Treatment of Hemorrhage or Hematoma	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.14	*	20200424	04/24/2020	Published
Comorbidity Risk Factors	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.13	*	20200201	02/01/2020	Published
Opioid (Oral Tablets and Patches only)	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.12	*	20191220	12/20/2019	Published
Fracture Exclusions For Hip And Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.2	*	20190622	06/22/2019	Published
Mechanical Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.1	*	20190618	06/18/2019	Published
Sepsis Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.4	*	20190618	06/18/2019	Published

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Malignant Neoplasm Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.7	*	20190618	06/18/2019	Published
Pneumonia Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.6	*	20190618	06/18/2019	Published
Pulmonary Embolism Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.3	*	20190618	06/18/2019	Published
Procedures Resulted From Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.11	*	20190615	06/15/2019	Published
Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.10	*	20190615	06/15/2019	Published
Procedures Resulted From Periprosthetic Joint Infection/Wound Infections	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.9	*	20190614	06/14/2019	Published
Periprosthetic Joint Infection/Wound Infection and Other Wound Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.8	*	20190614	06/14/2019	Published
Nonprimary Total Hip, Total Knee Replacement	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.5	*	20190613	06/13/2019	Published

***This cell is intentionally left empty**

Risk-Adjustment:

- **Issue 1:** Concern surrounding rationale for risk-adjustment of social variables including race.
    - **Developer response 1:** This measure includes risk adjustment for social variables, including African American race. Race was included in the risk-adjustment model to account for disparate rates of adverse events seen in literature (Stone et al., 2019), as well as to account to external stressors that impact health outside of healthcare provided to African American patients (Ard & Bullock, 2020; Ohm 2019). There is concern that risk-adjustment for race in measures designed for use in payment programs lets physicians “off the hook” for worsened rates in minority patient groups, while the counter argument sees the removal of risk-adjustment for race as a punitive measure against surgeons who perform arthroplasties on higher proportions of African American patients. The risk-adjustment model and rationale for inclusion of all variables

was analyzed at length in literature reviews and TEP panels and solicited feedback in qualitative interviews conducted with Massachusetts Health Quality Partners (MHQP) and in public comment periods throughout measure development. Ultimately, measure developers found that the benefits of including risk adjustment in the model outweighed the concerns; if implemented, we recommend ongoing monitoring of the impacts of the risk-adjustment model on provider-group payment.

  - Stone AH, et al. (2019). "Differences in perioperative outcomes and complications between African American and white patients after total joint arthroplasty". Journal of Arthroplasty 34(4):656-662
  - Ard K, Bullock C. (2020). "Concentrating risk? The geographic concentration of health risk from industrial air toxics across America. In Spatiotemporal Analysis of Air Pollution and Its Application in Public Health." (pp. 277-292). Elsevier.
  - Ohm, J.E. (2019). "Environmental exposures, the epigenome, and African American women's health." Journal of Urban Health, 96(1), pp.50-56.

- **Issue 2:** Information in ridge regression approach.
  - **Developer response 2:** We performed a ridge regression for colinear covariates, which allowed us to increase the predictive ability of the model and include 29 comorbid conditions (using ICD10 codes) within the model. This approach allowed for more variables to be included than a typical stepwise regression. All risk factors that were tested were included in the final model except for conditions that did not occur in the sample population. A full list of all comorbid conditions included in the model and the associated logistic regression coefficient estimates have been provided within the MIMS Intent to Submit application.

The BWH development team would like to thank reviewers and the NQF Methods Panel for their constructive feedback and support during the Intent to Submit process.

Measure Number: 3650e

Measure Title: Risk-standardized inpatient respiratory depression (IRD) rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM

Measure Developer/Steward: Brigham and Women's Hospital

Reliability

- **Issue 1:** Concern that the eCQM has only been tested in two EHRs from large, academic medical centers.
  - **Developer Response 1:** We tested this measure across 17 clinician groups in two geographically different healthcare systems with commonly used EHR vendors, Epic and Cerner. Measure testing in additional sites is costly and burdensome, the sample size and homogeneity of patients is noted as a limitation of this measure. We do not make the claim that this measure is generalizable to the larger population. Future analysis with additional healthcare systems is expected to increase variation in risk adjusted IRD rates.
- **Issue 2:** Concern over low ICC (ICC=0.069).
  - **Developer Response 2:** The ICC of 0.069 reflects the variation in patient score attributed to providers, a measure which tends to be <5%. This low ICC is likely attributed to the

small sample size; inclusion of additional clinician groups in analysis is expected to raise the ICC and decrease the range of the confidence intervals.
    - A separate ICC was calculated to describe how much variation in the provider-level scores is due to provider-level signal variation in the 2019 sample, resulting in an ICC of 0.972, well above the NQF required 0.50. We chose to only include 2019 in this ICC to demonstrate how the measure can meaningfully be used to report data on an annual basis. An ICC with all four years of data collected is expected to be higher.
- **Issue 3:** Concerns over lack of variation in rates across sites.
    - **Developer Response 3:** This measure has been tested in two geographically different sites, with risk-adjusted rates ranging from 1.92%-3.68%. Although variation across the sample is small, testing in two large, well performing medical centers still points to room for local quality improvement. We believe that a larger range in rates may be seen with more healthcare systems included in analysis.

Validity

- **Issue 1:** Concerns over providers' influence on IRD outcomes, potential meaningfulness of the measure as specified.
    - **Developer Response 1:** This measure was designed under cooperative agreement with CMS, where the measure was designed for use within the Merit-based Incentive Payment System (MIPS) which reports at the clinician group level. A limitation of reporting at the clinician group level is that respiratory depression can be a multidisciplinary issue where multiple disciplines including orthopedics, anesthesia and nursing among others may be responsible when this complication occurs. Given the broad range of providers involved in the prevention and management of inpatient respiratory depression, it may be more meaningful to report this measure at the hospital or health system level. During measure development, our TEP advised that the measure be reported as a facility-based measure. Facility based measures would be reported similarly to the overall rates within sites 1 and 2, rather than by clinician group. This demonstrates that the eCQM is already capable of capturing this level of analysis. Although we believe that this measure is meaningful at the clinician group level, we can see the benefits of transitioning this measure to report at the facility level. We will defer this decision to the NQF methods panel to assess which reporting level they believe could be most meaningful to healthcare systems and patients.
- **Issue 2:** Potential for random measurement error in SpO2 levels, concern over resulting unintended consequences.
    - **Developer Response 2:** According to literature (Ayad, Iqbal, & Singla, 2019), 90% oxygen saturation is considered respiratory depression. We moved this threshold down to 88% to be considered respiratory depression in the IRD measure as this more conservative estimate prevents overestimation resulting from random measurement error. We believe that it is better to miss a few cases rather than over penalize providers due to errors in measurement. This perspective was supported by our TEP.
    - We agree that there is a risk that continuous monitoring will flag more patients for IRD, but different safeguards can be put in place to ensure measure validity. For example, at

least two SpO2 readings are required for inclusion in the numerator. The move towards continuous monitoring promotes patient safety.

- S. Ayad, A.K. Khanna, S.U. Iqbal, N. Singla, Characterization and monitoring of postoperative respiratory depression: current approaches and future considerations, Br J Anaesth 123 (2019), 378-391.

- **Issue 3:** Concern over face validity vote (3/7).
  - **Developer Response 3:** Several our TEP members were concerned with the level of analysis of the IRD eCQM and believed that this measure would be more meaningful as a facility level measure, rather than a clinician group level measure. We believe that if our TEP were to vote on this measure again, as a facility level measure, that we would receive a higher level of face validity. As noted above, we will defer the decision regarding the level of analysis for this measure to the NQF methods panel.

Other General Comments

Measure Specifications:

- **Issue 1:** SNOMED CT codes [pull from RSCR].
  - **Developer Response 1:** The codes sets for all of our measures are published in the Value Set Authority Center (VSAC) and use standard codes that are recommended by CMS Blueprint and Promoting Interoperability program (see table below displaying query from NLM's VSAC):

Table 1: Brigham and Women's Hospital Published Code Sets

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Opioid Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.46	*	20210723	07/23/2021	Published
Total Knee Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.33	*	20210311	03/11/2021	Published
Total Hip Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.32	*	20210311	03/11/2021	Published
Active Bleeding	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.28	*	20201122	11/22/2020	Published
Anticoagulant Medications, Oral	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.20	*	20200804	08/04/2020	Published
Anticoagulant Medications, Injection	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.21	*	20200804	08/04/2020	Published
Anticoagulant Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.19	*	20200714	07/14/2020	Published
Acute Respiratory Failure	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.18	*	20200708	07/09/2020	Published
Coagulation Disorder Conditions	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.17	*	20200428	04/28/2020	Published
General Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.15	*	20200424	04/24/2020	Published
Treatment of Hemorrhage or Hematoma	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.14	*	20200424	04/24/2020	Published
Comorbidity Risk Factors	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.13	*	20200201	02/01/2020	Published
Opioid (Oral Tablets and Patches only)	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.12	*	20191220	12/20/2019	Published

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Fracture Exclusions For Hip And Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.2	*	20190622	06/22/2019	Published
Mechanical Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.1	*	20190618	06/18/2019	Published
Sepsis Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.4	*	20190618	06/18/2019	Published
Malignant Neoplasm Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.7	*	20190618	06/18/2019	Published
Pneumonia Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.6	*	20190618	06/18/2019	Published
Pulmonary Embolism Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.3	*	20190618	06/18/2019	Published
Procedures Resulted From Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.11	*	20190615	06/15/2019	Published
Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.10	*	20190615	06/15/2019	Published
Procedures Resulted From Periprosthetic Joint Infection/Wound Infections	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.9	*	20190614	06/14/2019	Published
Periprosthetic Joint Infection/Wound Infection and Other Wound Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.8	*	20190614	06/14/2019	Published
Nonprimary Total Hip, Total Knee Replacement	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.5	*	20190613	06/13/2019	Published

***This cell is intentionally left empty.**

Risk Adjustment:

- **Issue 1:** Rationale for risk adjustment of social variables including race.
    - **Developer Response 1:** This measure includes risk adjustment for social variables, including African American race. Race was included in the risk-adjustment model to account for disparate rates of adverse events seen in literature (Stone et al., 2019), as well as to account to external stressors that impact health outside of healthcare provided to African American patients (Ard & Bullock, 2020; Ohm 2019). There is concern that risk-adjustment for race in measures designed for use in payment programs lets physicians "off the hook" for worsened rates in minority patient groups, while the counter argument sees the removal of risk-adjustment for race as a punitive measure against surgeons who perform arthroplasties on higher proportions of African American patients. The risk-adjustment model and rationale for inclusion of all variables was analyzed at length in literature reviews and TEP panels. We also solicited feedback in qualitative interviews conducted with Massachusetts Health Quality Partners (MHQP) and in public comment periods throughout measure development. Ultimately, measure developers found that the benefits of including risk adjustment in the model outweighed the concerns; if implemented, we recommend ongoing monitoring of the impacts of the risk-adjustment model on provider-group payment.
    - The CMS Blueprint sees risk-adjustment in outcome measures as acceptable, however, this measure would be entirely functional in the EHR as an unadjusted eCQM. While as developers and our TEP support the use of risk-adjustment in this eCQM, we defer to NQF for the final decision of if this measure should move forward with or without the inclusion of a risk-adjustment model.
        - Stone AH, et al. (2019). "Differences in perioperative outcomes and complications between African American and white patients after total joint arthroplasty". Journal of Arthroplasty 34(4):656-662
        - Ard K, Bullock C. (2020). "Concentrating risk? The geographic concentration of health risk from industrial air toxics across America. In Spatiotemporal Analysis of Air Pollution and Its Application in Public Health." (pp. 277-292). Elsevier.
        - Ohm, J.E. (2019). "Environmental exposures, the epigenome, and African American women's health." Journal of Urban Health, 96(1), pp.50-56.

The BWH development team would like to thank reviewers and the NQF Methods Panel for their constructive feedback and support during the Intent to Submit process.

Measure Number: 3652e

Measure Title: Risk-standardized prolonged opioid prescribing rate following elective primary total hip arthroplasty (THA) and/or total knee arthroplasty (TKA) eCQM

Measure Developer/Steward: Brigham and Women's Hospital

Reliability

- **Issue 1:** Clarification on measure type.
    - **Developer Response 1:** The risk-standardized prolonged opioid prescribing rate is a process measure eCQM. This measure is not an outcome measure, nor is it a PRO-PM.
- **Issue 2:** Concern regarding the limited number of sites included in testing and variation in rates.
    - **Developer Response 2:** We tested this measure across 13 clinician groups in two geographically different healthcare systems with commonly used vendors, Epic and Cerner. Measure testing in additional sites is costly and burdensome, the sample size

and homogeneity of patients is noted as a limitation of this measure. Regarding variation in rates across THA and TKA samples (THA: 2.48%-23.53%; TKA: 5.38%-36.08), we found differences across both clinician groups and healthcare systems which points to room for quality improvement following measure implementation.

- **Issue 3:** Concern over low ICCs in THA and TKA samples.
    - **Developer Response 3:** The ICCs of 0.0929 (THA), 0.11675 (TKA) reflect the variation in patient score attributed to providers, a measure which tends to be <5%. The ICCs are likely attributed to the small sample analyzed; inclusion of additional clinician groups in analysis is expected to raise the ICCs and decrease the range of confidence intervals.
    - Separate ICCs were calculated to describe how much variation in the provider-level scores is due to provider-level signal variation in the 2019 sample, resulting in ICCs of 0.9566 (THA) and 0.9691 (TKA), well above the NQF required 0.50. We chose to only include 2019 data in this ICC to demonstrate how the measure can meaningfully be used to report data on an annual basis. An ICC with all four years of data collected is expected to be higher.

Validity

- **Issue 1:** The process of correctly abstracting information the EHR will need to be repeated with each future EHR. The authors seem to be showing that the measure can be made valid, not that it is valid 'off the shelf.'
    - **Developer Response 1:** This is not accurate. eCQM refinement was part of our development process. We conducted multiple rounds of chart reviews in defining measure specifications and in building an eCQM that can accurately capture codes. Following measure development and refinement through this iterative process in MGB testing, we worked with Cerner to implement the eCQM in Site 2. Additional tweaks were not required to implement at the Cerner site. Now that the measure has been defined, it is not expected to require adaptions or tweaks in future EHR implementation.
- **Issue 2:** Wide confidence intervals around the measure at the site level.
    - **Developer Response 2:** Wide confidence intervals are likely caused by the small sample used in analysis; we expect the range of confidence intervals to decrease with the inclusion of additional sites in analysis.
- **Issue 3:** Concern that validity was only assessed as face validity by a TEP
    - **Developer Response 3:** Our TEP was comprised of orthopedic surgeons, clinicians, patient representatives, and measure development professionals. Although the face validity vote itself was comprised of a single question ("*The performance scores resulting from the risk-standardized prolonged opioid prescribing rate eCQM, as specified, can be used to distinguish good from poor clinician group-level quality related to patient safety*."), it is backed by continuous meetings with the TEP throughout three years of measure development where the TEP had an opportunity to discuss ongoing development of measure specifications and analysis, provide feedback, and ask questions.
    - Chart reviews to assess agreement between manual reviewer and the eCQM are typically seen as validity testing, however, the BWH team chose not to label either the manual chart review or the eCQM output as the "gold" standard as both methods

demonstrated room for inaccuracies during the testing process. The BWH team has tested the data abstraction accuracy of the eCQM, and we have shown that it reliably pulls the correct information from the EHR. We acknowledge that the eCQM relies on the correct input of information into the EHR. Conversely, manual chart review and the ability to read through progress notes inherently provides more qualitative information than what can be captured by an eCQM. However, with manual review of the EHR notes, at times BWH reviewers would incorrectly interpret information as procedures/diagnostic codes are occasionally buried or hidden within the EHR. Due to the drawbacks of both methods, the BWH team decided that reporting a kappa value evaluating the similarity of results between both methods would be more appropriate than generating a PPV and/or NPV.

Other General Comments

Measure Specifications:

- **Issue 1:** SNOMED-CT codes/ICD10 codes
  - **Developer Response 1:** The codes sets for all of our measures are published in the Value Set Authority Center (VSAC) and use standard codes that are recommended by CMS Blueprint and Promoting Interoperability program (see table 1 below displaying query from NLM's VSAC):

Table 1: Brigham and Women's Hospital Published Code Sets

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Opioid Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.46	*	20210723	07/23/2021	Published
Total Knee Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.33	*	20210311	03/11/2021	Published
Total Hip Arthroplasty Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.32	*	20210311	03/11/2021	Published
Active Bleeding	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.28	*	20201122	11/22/2020	Published
Anticoagulant Medications, Oral	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.20	*	20200804	08/04/2020	Published
Anticoagulant Medications, Injection	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.21	*	20200804	08/04/2020	Published
Anticoagulant Medications	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.19	*	20200714	07/14/2020	Published
Acute Respiratory Failure	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.18	*	20200708	07/09/2020	Published
Coagulation Disorder Conditions	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.17	*	20200428	04/28/2020	Published
General Surgery	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.15	*	20200424	04/24/2020	Published
Treatment of Hemorrhage or Hematoma	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.14	*	20200424	04/24/2020	Published
Comorbidity Risk Factors	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.13	*	20200201	02/01/2020	Published
Opioid (Oral Tablets and Patches only)	Extensional	RXNORM	BWH	BWH	2.16.840.1.113762.1.4.1206.12	*	20191220	12/20/2019	Published

Name	Type	Code System	Steward	Author	OID	Keyword	Latest Version	Updated Date	Status
Fracture Exclusions For Hip And Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.2	*	20190622	06/22/2019	Published
Mechanical Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.1	*	20190618	06/18/2019	Published
Sepsis Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.4	*	20190618	06/18/2019	Published
Malignant Neoplasm Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.7	*	20190618	06/18/2019	Published
Pneumonia Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.6	*	20190618	06/18/2019	Published
Pulmonary Embolism Complications Related To Hip and Knee Procedures	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.3	*	20190618	06/18/2019	Published
Procedures Resulted From Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.11	*	20190615	06/15/2019	Published
Surgical Site Bleeding and Other Surgical Site Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.10	*	20190615	06/15/2019	Published
Procedures Resulted From Periprosthetic Joint Infection/Wound Infections	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.9	*	20190614	06/14/2019	Published
Periprosthetic Joint Infection/Wound Infection and Other Wound Complications	Extensional	ICD10CM	BWH	BWH	2.16.840.1.113762.1.4.1206.8	*	20190614	06/14/2019	Published
Nonprimary Total Hip, Total Knee Replacement	Extensional	ICD10PCS	BWH	BWH	2.16.840.1.113762.1.4.1206.5	*	20190613	06/13/2019	Published

***This cell is intentionally left empty.**

- **Issue 2:** Basis for the interval of ">42 days" following the THA/TKA procedure.
    - **Developer Response 2:** Per the Washington State Guideline used to guide measure development, opioids ideally should be tapered off within 14 days of THA/TKA except in

exceptional cases, which should be tapered off within 42 days. The use of this guideline to define prolonged opioid prescribing was approved by our TEP.

- Washington state agency medical directors' Group (2018). Interagency guidelines on prescribing opioids for pain. Available at http://www.agencymeddirectors.wa.gov/Files/FinalSupBreeAMDGPostopPain091318wcover.pdf

Risk Adjustment:

- **Issue 1:** Appropriateness of the calibration and discrimination of the models.
  - **Developer Response 1:** We performed a ridge regression for colinear covariates, which allowed us to increase the predictive ability of the model and include 29 comorbid conditions (using ICD10 codes) within the model. This approach allowed for more variables to be included than a typical stepwise regression. All risk factors that were tested were included in the final model except for conditions that did not occur in the sample population. A full list of all comorbid conditions included in the model and the associated logistic regression coefficient estimates have been provided within the MIMS Intent to Submit application.
  - **Issue 2:** Rationale for risk-adjusting for social factors, including race.**Developer Response 2:** This measure includes risk adjustment for social variables, including African American race. Race was included in the risk-adjustment model to account for disparate rates of adverse events seen in literature (Stone et al., 2019), as well as to account for external stressors that impact health outside of healthcare provided to African American patients (Ard & Bullock, 2020; Ohm 2019). There is concern that risk-adjustment for race in measures designed for use in payment programs lets physicians "off the hook" for worsened rates in minority patient groups, while the counter argument sees the removal of risk-adjustment for race as a punitive measure against surgeons who perform arthroplasties on higher proportions of African American patients. The risk-adjustment model and rationale for inclusion of all variables was analyzed at length in literature reviews and TEP panels. We also solicited feedback in qualitative interviews conducted with Massachusetts Health Quality Partners (MHQP) and in public comment periods throughout measure development. Ultimately, measure developers found that the benefits of including risk adjustment in the model outweighed the concerns; if implemented, we recommend ongoing monitoring of the impacts of the risk-adjustment model on provider-group payment.
  - The CMS Blueprint sees risk-adjustment in process measures as acceptable, however, this measure would be entirely functional in the EHR as an unadjusted eCQM. While as developers support the use of risk-adjustment in this eCQM, we defer to NQF for the final decision of if this measure should move forward with or without a risk-adjustment model (we provided that data analysis for both in our submission).
    - Stone AH, et al. (2019). "Differences in perioperative outcomes and complications between African American and white patients after total joint arthroplasty". Journal of Arthroplasty 34(4):656-662

- Ard K, Bullock C. (2020). “Concentrating risk? The geographic concentration of health risk from industrial air toxics across America. In Spatiotemporal Analysis of Air Pollution and Its Application in Public Health.” (pp. 277-292). Elsevier.
- Ohm, J.E. (2019). “Environmental exposures, the epigenome, and African American women’s health.” Journal of Urban Health, 96(1), pp.50-56.

The BWH development team would like to thank reviewers and the NQF Methods Panel for their constructive feedback and support during the Intent to Submit process.

Measure Number: 3638

Measure Title: Care Goal Achievement Following a Total Hip Arthroplasty (THA) or Total Knee Arthroplasty (TKA)

Measure Developer/Steward: Brigham and Women’s Hospital

Developer Opening Comments

We would like to thank all the reviewers for their constructive comments. We found them very helpful.

We appreciate the varied feedback of reviewers regarding our testing methodology and outcomes, and the acknowledgement of our reliability and validity methodologies and their limitations. The reviewers provided detailed feedback on various issues surrounding the outcomes of the reliability and validity testing. We have taken seriously each issue raised and provided a detailed response accordingly.

We would like to clarify that the patient reported outcome performance measure (PRO-PM) is the measure being submitted for endorsement, not the patient reported outcome measures (PROMs) that were submitted as an executive summary attachment. We developed the PROMs first, and then developed and tested the PRO-PM. Therefore, while the context of the PROMs becomes relevant information to review when looking at the PRO-PM submission, the evaluation and decisions should be related to the new PRO-PM.

We feel it is important to mention this because when we reviewed the specific comments, we noticed that some of the feedback (i.e., written and checkboxes) was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. For example, some feedback mentioned that we had tested the PRO-PM on the Group/Practice level as specified, yet other feedback (e.g., Reviewer #3) noted that the measure was only tested at the patient level.

Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality it had been conducted and completed on the appropriate level, which is Group/Practice level. Importantly, although the PROMs and PRO-PM are related, we used different testing methodologies when we tested these measures, which led to different specifications and outcomes. Therefore, we do have some worries about some of the reviewer’s preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes.

Regarding the issues related to small sample size and low variability, which were raised by some of the reviewers, we do believe that the outcome of the reliability and validity testing of the proposed PRO-PM should be assessed in the context of a newly developed PROMs, not based on already established PROMs widely used in clinical settings and/or that are part of an existing data collection registry. While the development of other PRO-PMs based on existing PROMs whereby there is already large volumes of legacy data to utilize could be tested with a large sample size and diverse clinician groups, our PRO-PM development constituted of prospective data collection and analysis in real-time, real workflow settings, which in our case required collecting paired survey patient data before and after total hip or total knee replacement surgery. Therefore, while we acknowledge the issues related to small sample size and low variability resulting in inconclusive results on reliability and validity, we do feel that testing this new PRO-PM with a larger sample size and other clinician groups would delay the use of this PRO-PM by many years. Based on our comprehensive qualitative interviews with patients, providers, and payers, along with environmental scans, we believe that there is a real need and a great value proposition for a performance measure based on the concept of care goal achievement in total hip arthroplasty and total knee arthroplasty, and that this domain is not measured currently by an existing measure in widespread use.

Our measure is designed to promote patient-centered care and enable care that is personalized and aligned with patient's goals, and more importantly, it fills a void in the PRO-PM development realm. Therefore, we kindly ask that this key factor be taken into consideration during the review process.

As mentioned, we have taken each issue raised by the reviewers very seriously and have provided a detailed response accordingly. We hope that our responses have provided appropriate clarification for the issues that reviewers mentioned. If needed, we will be happy to present and provide more information to the Scientific Methods Panel.

We thank the Committee and its reviewers for their helpful feedback and for considering our measure.

Reliability

Note 1: We addressed each issue identified separately. In some instances, when there have been multiple reviewers raising the same issue(s)/concern(s), we consolidated them to provide one response, and in other instances, we have addressed issues on an individual basis.

Note 2: For ease of readability and specificity, we included the question number noted in the NQF- 3638 SMP SA PA Form_Combined-508 document we received which details the reviewers' feedback. We have also included the specific reviewer where possible.

Issue 1: Concerns about the measure specifications - In question 2, reviewer 2 questioned if the variability of the interval length across patients could introduce bias in achievement of goals and if there is empirical evidence to show that it doesn't.

- **Developer Response 1:** Thank you for taking the time to review our submission and for pointing this specific issue. The variability of time for data collection was structured purposely to reflect the most suitable timeframe allow to capture goals achievements following total hip arthroplasty (THA) and total knee arthroplasty (TKA) and minimize potential biases. Based on the mixed methods below we defined the timeframes of the new measure and minimized the potential biases related to the variability of the interval length across patients.

The measure timeframes were defined based on comprehensive qualitative and quantitative testing methods. Specifically, we obtained key stakeholders' input at several points during the measure development process. This qualitative approach consisted of semi-structured interviews and focus groups with patients, health care providers, payers and experts in the field about the measure specification, including the timeframes. These stakeholders supported the proposed measure timeframes and did not raise any concerns related to biases.
The measure specifications, including the timeframes, were also assessed and defined based on research on relevant measures to improve alignment (i.e., measure harmonization) across various other PROMs and other measures in the MUC List and NQF database related to orthopedic surgery. Consistent with this notion, other relevant measures specified for similar timeframes. In addition, the variability of time for data collection was also structured so as to not create bias given that THA and TKA patients heal at different times – the timeframes allow for this differentiation and the analyses were stratified as a result.
Finally, the measure's timeframes were also defined based on quantitative testing methods (i.e., retrospective and prospective data analysis). The quantitative testing included analysis of existing PROM datasets (e.g., PROMIS 10, HOOS-PS and KOOS-PS) within the Mass General Brigham (MGB) Enterprise Data Warehouse (EDW) completed by total hip and total knee replacement patients... The care goal achievement (CGA) surveys were incorporated into MGB's Musculoskeletal (MSK) questionnaire set that is assigned to THA and TKA patients before and after surgery and administered through MGB's electronic survey platform. Thus, we were able to test our measure specifications, including various timeframes and define it based on these quantitative testing methods. It is important to mention that potential bias related to the variability of the interval length across patients was not directly empirically assessed.
Using the mixed methods above, we defined the most suitable timeframes for the new measure, including minimizing the potential bias related to the variability of the interval length across patients. Importantly, throughout the measure development process, the care goal achievement PRO-PM and its timeframes, were vetted by our TEP members, patient representatives, orthopedic surgeons specializing in THA and TKA, and measure developers. These key stakeholders supported the use of the proposed timeframes.

- **Issue 2: Concerns about the measure specifications –** In question 2, reviewer 5 noted that in the data dictionary, CPT code 27445 is a denominator inclusion code and is listed as such under the question sp15 in the "form information," yet is also listed in the exclusion set of codes under the rubric of "revision procedure of the THA or TKA."
    - **Developer Response 2:** Thank you for taking the time to review our submission. We appreciate and acknowledge your concern. The CPT code 27445 is listed in both places because when used without a modifier code, it is a total knee replacement which is included as our measure specification inclusion criteria. When CPT 27445 is used with a modifier code of 78 (27445-78), it becomes specified as a revision of a total knee replacement, which is part of our measure specification exclusion criteria and is noted as such in the data dictionary on the revision tab.
- **Issue 3: Concerns about the measure specifications –** In question 2, reviewer 11 noted that the measure specifications were clear and adequate but had some concerns about checking the completeness of the data collected. For instance, how would someone know if the surveys were being given only to those patients likely to do well or if there was a bias in response among those invited to participate.

- **Developer Response 3:** Thank you for taking the time to review our submission. We appreciate your comment and concern. Our patient reported outcome measure (PROM) surveys, from which the patient reported outcome performance measure (PRO-PM) is derived, are appended to an already established survey set that is administered to all patients scheduled for a total hip arthroplasty (THA) or total knee arthroplasty (TKA) in all six testing sites (clinician-groups).
  The measure development team was able to operationalize the testing of the PROM surveys, in a real use case scenario, using the clinician group's electronic health record (EHR) system. All the testing sites (clinician-groups) were already utilizing EHR to capture patient-reported outcomes (PROs) from patient's using the platform's data collection functionality in clinic via iPad and at home via patient portal. Both the newly developed pre- and post- surgical surveys were built, programmed, and appended to the sites Musculoskeletal Questionnaire Set which also included other PROs, such as PROMIS-10, HOOS-PS, and KOOS-PS, to collect survey data from surgical patients.
  Our implementation mirrors the workflow of many orthopedic practices utilizing PROMs whereby the administration of the measure is executed as an automated process, i.e., linked to a surgery or appointment. Consistent with this approach, we were able to assess the completeness of the data collected, as well as to minimized and possibly eliminate the option that the surveys were being given only to those patients likely to do well. While it is possible that a practice could administer the surveys to only those patients they feel would do well, the burden on the work staff to initiate such a process would be cumbersome and costly. Also given the patient case numbers needed to submit for reimbursement (n=25), most likely the 'N' of an invitation-only group would not be sufficient for submission.

- **Issue 4: Data Source (checkboxes) –** In the Reliability Testing section (Page 3), reviewer 4 selected 'Other (please specify) checkbox and wrote in 'Paper' as a Data Source .
  - **Developer Response 4:** Thank you for taking the time to review our submission and your suggestion to add also 'Paper' as a Data Source. We are unclear as to what the write-in statement refers to as our survey was administered only electronically during all phases of testing. Could you please provide more clarification as we can address your suggestion/concern?
- **Issue 5: Level of Analysis selection (checkboxes) –** In the Reliability Testing/Level of Analysis section (Pages 3&4), reviewer 3 noted that although developers indicated Group/Practice level analyses, the reviewer has identified only patient level analyses.
  - **Developer Response 5:** Thank you for taking the time to review our submission and for pointing this specific issue. For the level of analysis done for the PRO-PM, we selected 'Group/Practice' under the Reliability Testing section as specified, however, reviewer 3 noted that only patient-level analyses had been completed. We would like to clarify that for the PRO-PM we conducted and completed our testing on the appropriate level, which is Group/Practice level. Our Group/Practice level analyses is described in the formal electronic measure submission.

    In our submission, in order to set the context for the PRO-PM, we included a file titled "PROM Testing Executive Summary," which explained the reliability and validity of the PROMs used to develop the PRO-PM. As mentioned in the opening comments, although the PROMs and PRO-PM are related, we used different testing methodologies

when we tested these measures, which led to different specifications and outcomes. We are concerned that reviewer 3 may have reviewed the PROM supplement only, not our main application regarding the PRO-PM. We say this because reviewer 3 did not reference any of the analyses presented in the PRO-PM document and even mentioned not being able to find any analyses at all on the Group/Practice level.

Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality it had been conducted and completed on the appropriate level, which is Group/Practice level. Consistently, we do have some worries about the reviewer's preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes and consequently, the overall rating of the reliability and validity. Please let us know what you recommend in this circumstance.

- **Issue 6: The sample of clinician groups is small –** In the sections on reliability methods, results, and rational for overall rating (questions 6, 7, and 11) most of the reviewers made comments related to the small sample of clinician groups**.** The reviewers frequently raised the issue of the small number of clinician groups in our PRO-PM development sample and the resulting methodological difficulties for both reliability and validity testing.
    - **Developer Response 6:** We thank the reviewers for their constructive comments related to the small sample of clinician groups. While we acknowledge the issues related to the small sample size, we would like to put forward some general thoughts.
      The Care Goal Achievement (CGA) PRO-PM is an entirely new measure, based on a new PROMs developed for this purpose. We believe that the outcome of the reliability and validity testing of the proposed PRO-PM should be assessed in the context of a newly developed PROMs, not based on already established PROMs widely used in clinical settings and/or that are part of an existing data collection registry. While the development of other PRO-PMs based on existing PROMs whereby there is already large volumes of legacy data to utilize could be tested with a large sample size and diverse clinician groups, our PRO-PM development constituted of **prospective data collection** and analysis in real-time, real workflow settings, which in our case required collecting paired survey patient data before and after total hip or total knee replacement surgery. Taking in consideration the limitation of the small sample size, it is important to mention that we tested our PRO-PM in 6 sites (clinician groups). As stated in another section, we were able to operationalize the testing of the PROM surveys, in a real use case scenario, using the clinician group's electronic health record (EHR) system. Both the newly developed pre- and post- surgical surveys were built, programmed, and appended to the sites Musculoskeletal Questionnaire Set which also included other PROs, such as PROMIS-10, HOOS-PS, and KOOS-PS, to collect survey data from surgical patients. As a results, our implementation and testing mirrors the workflow of many orthopedic practices utilizing PROMs, which confirms the suitability, feasibility and usability of our new PRO-PM.
      While we acknowledge the issues related to small sample size, we do feel that testing this new PRO-PM with a larger sample size and other clinician groups would delay the use of this PRO-PM by many years. Based on our comprehensive qualitative interviews with patients, providers, and payers, along with environmental scans, we determined that there is a real need and a great value proposition for a performance measure based

on the concept of care goal achievement in total hip arthroplasty and total knee arthroplasty, and that this domain is not measured currently by an existing measure in widespread use. Consistent with this notion, our measure is designed to promote patient-centered care and enable care that is personalized and aligned with patient's goals, and more importantly, it fills a void in the PRO-PM development realm.
In summary, while we understand the reluctance to endorse a measure that were tested on a small sample size, we kindly ask that these key factors be taken into consideration during the review process of our innovative and valuable PRO-PM. If needed, we will be happy to present and provide more information to the Scientific Methods Panel.

- **Issue 7: Low variability due to small sample size –** Three reviewers commented on the lack of variation in scores between practice groups, raising concerns about the ability of the PRO-PM to distinguish between practices (reviewer 2 and 7, question 7; reviewer 11, question 11).
    - **Developer Response 7:** We thank the reviewers for their constructive comments related to the low variability in scores between practice groups. We recognize the low variability in group level scores but believe this is a function of the small sample size as well as the homogeneity of the practices that were tested. For additional information related to the small sample size issue, please review our previous response (Issue number 6). The homogeneity of the practices might be related to the fact that all the **6** testing sites (hospitals/clinician groups) are affiliated to the same health system.
    Importantly, the value of the measure and its ability to distinguish between practices was also assessed with stakeholders in qualitative assessment (i.e., interviews and focus groups) throughout the measure development process. Patients and providers saw great value in the new PRO-PM and its scoring. Providers thought that the PRO-PM will be a very good tool to assess the performance of clinician groups and compare their outcomes related to care goal achievement. Payers' interviews also supported these findings and added that the new PRO-PM will enable a new national benchmark related to care goal achievement and possibly incentivize efforts to implement the necessary improvements to practice quality.
    Based on the above information, we believe that the PRO-PM may well distinguish between practices, especially between heterogeneous practices; it just needs the opportunity for such wider scale implementation.
- **Issue 8: The sample of clinician groups is too small to conduct the boot strapping procedure –** Reviewer 2 (question 7) mentioned that the sample size was too small to conduct the boot strapping procedure in particular.
    - **Developer Response 8:** We appreciate and acknowledge this perspective. We carried out the bootstrapping process to attempt to address the questions raised by the small sample and agree that it did not result in a different conclusion. We included the analysis for completeness.
- **Issue 9: Threshold for internal consistency reliability is too low –** In questions 6 and 7, Reviewer 3 discussed the threshold we used for evaluating the internal consistency reliability (Cronbach's alpha) of the pre- and post-surgical Care Goal Achievement (CGA) PROMs.
    - **Developer Response 9:** We appreciate your comment and concern. This issue brought up was specific to the CGA PROMs measure development, not to the PRO-PM development. In questions 6 and 7, Reviewer 3 discussed the threshold we used for evaluating the internal consistency reliability (Cronbach's alpha) of the pre- and post-surgical CGA PROMs. While we set a threshold of 0.7, the reviewer suggested 0.9 as more appropriate. We would be happy to debate about the threshold but would like to put the question in a bit of perspective. Cronbach's alpha is a measure of a

psychometric scale, which in our case consisted of a continuous score constructed from the 8 items on the tool. We presented this in the PROM supplement to show the development of the PROMs as standalone instruments. However, for the PRO-PM application, we did not use these scales. Instead, we compared the patient's pre- and post-surgical responses on all 8 items and scored each as a binary: met expectations =1 / did not meet expectations = 0. If the patient met 6 or more expectations, then they were counted as a "pass" on the paired PROM, which was a data element for the PRO-PM. This binary outcome of "pass" or "not pass" on the patient level is actually the only measure that is relevant for the PRO-PM. So, the issue of the threshold on the PROM scales is not central to the use of the tools for the performance measure. For this reason, we would like to set aside the debate on Cronbach's alpha thresholds.
As we mentioned in one of the previous sections (Issue 5), we are concerned that reviewer 3 may have reviewed the PROM supplement only, not our main application regarding the PRO-PM. We say this because reviewer 3 did not reference any of the analyses presented in the PRO-PM document. Consistently, we do have some worries about the reviewer's preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes and consequently, the overall rating of the reliability and validity.

- **Issue 10: Unidimensionality is not established –** In question 7, reviewer 3 had concerns about our conclusion of unidimensionality for the PROM scores, given the factorial complexity of the EFA results.
    - **Developer Response 10**: Thank you for taking the time to review our submission and providing feedback about unidemensionality. Reviewer 3 (question 7) was concerned about our conclusion of unidimensionality for the PROM scores, given the factorial complexity of the EFA results. Our reasons for calling it unidimensional were listed in detail in our application.
      We would like to refer back to our response to issue 9 under reliability. The factor structure of the PROM instruments only relates to their use as a psychometric scale for patient assessment. But in the context of the CGA PRO-PM, the 8 items on the surveys are scored entirely differently, starting with a comparison of the pre- and post-responses and ending in a binary pass/fail for each patient. In this measurement scheme, the factor structure underlying the 8 items is not relevant.
      As we mentioned in the previous sections (Issue 5 & 9), we are concerned that reviewer 3 may have reviewed the PROM supplement only, not our main application regarding the PRO-PM. We say this because reviewer 3 did not reference any of the analyses presented in the PRO-PM document. Consistently, we do have some worries about the reviewer's preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes and consequently, the overall rating of the reliability and validity.
- **Issue 11: Reliability testing was conducted with the data source and level of analysis indicated for this measure (Question 4; checkboxes)**
    - **Developer Response 11:** We noticed that 'Yes' and 'No' were selected for question 4. While we acknowledge the opinions of the reviewers, we think that our reliability testing was conducted with the data source and level of analysis indicated for this PRO-PM. Thus, we provided appropriate reliability testing that matched the level of analysis as indicated in question 3.
      As we stated in the opening comments and other sections, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made

based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality, we conducted the appropriate testing for the PRO-PM, and it had been conducted and completed on the appropriate level.

- **Issue 12: Was the method described and appropriate for assessing the proportion of variability due to real differences among measured entities? (Question 8; checkboxes)**
    - **Developer Response 12:** We noticed that 'Yes', 'No' and 'Not applicable' were selected for question 8. While we acknowledge the various opinions of the reviewers, we think that the method we used was appropriate for assessing the proportion of variability due to real differences among measured entities. Please see our responses in issues 6 and 7 which address the appropriateness of the method used to assess the variability.
      As we stated in the opening comments and other sections, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM.
      Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality, we conducted the appropriate testing for the PRO-PM.
- **Issue 13: Was the method described and appropriate for assessing the reliability of ALL critical data elements? (Question 9; checkboxes)**
    - **Developer Response 13:** We noticed that 'Yes' and 'No' were selected for question 9. While we acknowledge the various opinions of the reviewers, we think that the method we used was appropriate for assessing the reliability of ALL critical data elements. Please see our responses in issues 6 and 7 which address the appropriateness of the method used to assess the variability.
      As we stated in the opening comments and other sections, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM.
      Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality, we conducted the appropriate testing for the PRO-PM.
- **Issue 14: Overall rating of reliability (Question 10; checkboxes)**
    - **Developer Response 14:** It looks like reviewers had different opinions regarding the overall rating of reliability. The various ratings that were selected included 'Moderate,' 'Low' and 'Insufficient.' While we acknowledge and appreciate the various opinions of the reviewers, we think that we used the appropriate reliability testing methods and provided clarifications to their comments and concerns. Please review our responses to the various issues raised by the reviewers in the reliability sections.
      As we stated in the opening comments and other sections, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. As a result, we do have some worries about the reviewer's preliminary

feedback about the appropriateness of our PRO-PM testing methodology and outcomes and consequently, the overall rating of the reliability.

Validity

Developers Note 1: We addressed each issue identified separately. In some instances, when there have been multiple reviewers raising the same issue(s)/concern(s), we consolidated them to provide one response, and in other instances, we have addressed issues on an individual basis.
Developers Note 2: For ease of readability and specificity, we included the question number noted in the NQF- 3638 SMP SA PA Form_Combined-508 document we received, which details the reviewers' feedback. We have also included the specific reviewer where possible.

Issue 1: The sample of clinician groups is small – In the sections on validity methods, results, and for overall rating (questions 16, 20, and 26), six reviewers (10, 7, 5, 9, 11, and 2) made comments related to the small sample of clinician groups**.** The reviewers frequently raised the issue of the small number of clinician groups in our PRO-PM development sample and the resulting methodological difficulties for validity testing.

- **Developer Response 1:** Thank you for taking the time to review our submission and raise these issues about small clinician group samples. While we acknowledge the issues related to the small sample size, we would like to put forward some general thoughts. The Care Goal Achievement (CGA) PRO-PM is an entirely new measure, based on a new PROMs developed for this purpose. We believe that the outcome of the reliability and validity testing of the proposed PRO-PM should be assessed in the context of a newly developed PROMs, not based on already established PROMs widely used in clinical settings and/or that are part of an existing data collection registry. While the development of other PRO-PMs based on existing PROMs whereby there is already large volumes of legacy data to utilize could be tested with a large sample size and diverse clinician groups, our PRO-PM development constituted of prospective data collection and analysis in real-time, real workflow settings, which in our case required collecting paired survey patient data before and after total hip or total knee replacement surgery. Taking in consideration the limitation of the small sample size, it is important to mention that we tested our PRO-PM in 6 sites (clinician groups). As stated in another section, we were able to operationalize the testing of the PROM surveys, in a real use case scenario, using the clinician group's electronic health record (EHR) system. Both the newly developed pre- and post- surgical surveys were built, programmed, and appended to the sites Musculoskeletal Questionnaire Set which also included other PROs, such as PROMIS-10, HOOS-PS, and KOOS-PS, to collect survey data from surgical patients. As a results, our implementation and testing mirrors the workflow of many orthopedic practices utilizing PROMs, which confirms the suitability, feasibility and usability of our new PRO-PM.
  While we acknowledge the issues related to small sample size, we do feel that testing this new PRO-PM with a larger sample size and other clinician groups would delay the use of this PRO-PM by many years. Based on our comprehensive qualitative interviews with patients, providers, and payers, along with environmental scans, we determine that there is a real need and a great value proposition for a performance measure based on the concept of care goal achievement in total hip arthroplasty and total knee arthroplasty, and that this domain is not measured currently by an existing measure in

widespread use. Consistent with this notion, our measure is designed to promote patient-centered care and enable care that is personalized and aligned with patient's goals, and more importantly, it fills a void in the PRO-PM development realm.
In summary, while we understand the reluctance to endorse a measure that were tested on a small sample size, we kindly ask that these key factors be taken into consideration during the review process of our innovative and valuable PRO-PM. If needed, we will be happy to present and provide more information to the Scientific Methods Panel.

- **Issue 2: Risk Adjustment –** In question 19e, several reviewers (2, 7, 9, and 11) noted collectively that the sample was homogeneous, and results of the risk modeling were not useful.
    - **Developer Response 2:** Thank you for taking the time to review our submission and raise these issues related to risk adjustment. Reviewer 5 felt our risk adjustment was reasonable and reviewer 11 also noted that the approach was acceptable. Reviewer 2 commented that the weights from the risk adjustment model were "based on a small, homogeneous patient sample and are unlikely to generalize." Reviewers 11 and 7 made similar observations in the same question. We recognized this issue in our application and explained that our choice of factors to adjust for was made conceptually, based on expert input, other relevant measures in orthopedics, and known factors from the literature. We made this decision conceptually in part because we knew our sample was homogeneous and not representative. We presented the risk adjustment results for completeness but noted how our adjusted scores were very close to the unadjusted scores, reflecting the nonsignificant model parameters.
    Reviewer 9 mentioned that risk adjustment may not be appropriate because it might mask important differences in the outcome. Similarly, reviewer 2 commented that "social determinants are very likely to influence patient expectations regarding outcomes of surgery and are a potential source of bias in the use of this measure as a quality of care indicator at the group level." We recognize the sometimes distorting effects of "adjusting away" important covariates when measuring quality, however, we believe that these concerns do not affect the care goal achievement (CGA) PRO-PM any more than any other performance measure. In addition, other existing and valid orthopedic measures use similar risk adjustment models.
    Finally, reviewer 2 noted that there was no validation of the risk model. We assume this refers to the class of cross-validation methods involving splitting the data into training and test datasets. While we would have liked to use this approach, we did not have the sample size to support it.
- **Issue 3: Face validity of measure score as an indicator of quality was not documented -** In question 16, reviewer 5 had concerns about evidence of face validity provided for the measure but not for the measure score as an indicator of quality.
    - **Developer Response 3:** Thank you for taking the time to review our submission and raise these issues related to face validity. Our application stated that our technical expertise panel (TEP) determined the face validity of *both* the measure itself and the measure score as an indicator of quality. Reviewer 5 (question 16) noticed that we provided evidence of face validity for the measure but not for the measure score as an indicator of quality. Throughout the various development stages of the measure specifications as well as after the measure was fully specified, a group of experts (n=6) from the TEP was assembled to vote on the face validity of the measure. The voting of the TEP provided face validity from a group of experts in the fields of orthopedic surgery, measure developers, and patient advocacy.
    A vote was conducted with the six members of the development team's TEP with 100 percent endorsing the validity of the measure. **Importantly**, **the TEP members also**

voted on the scoring of the measure, which included the face validity of a measure score as a quality indicator. TEP members acknowledged that the PRO-PM score, as specified, can be used to distinguish good care goal achievement or poor care goal achievement as an indicator of quality.

- **Issue 4: Handling of missing data –** In question 22, reviewers 2, 3, and 11 raised questions about missing data, how it is handled, and how audits are conducted.
    - **Developer Response 4:** Thank you for taking the time to review our submission and providing feedback how missing data handled. There were several reviewers who did not have any concerns about the missing data and 3 who wanted more clarification.
      We would like to draw the reviewers' attention to table 11 and section 2b.10 in our PRO-PM application, which detailed the extraordinarily low rate of missingness on the care goal achievement (CGA) survey level. There were 439 patients in the development sample, who each completed a pre-surgical and post-surgical PROM, each with 8 items. Of the 7,024 responses to these items in the data (439 x 2 timepoints x 8 items), only one item was missing, a missingness rate of 0.2%. Due to this low rate, we did not pursue further testing of survey missingness and the biases it can introduce.
      We attribute this low rate of missing data in part to the fact that our newly developed pre- and post- surgical surveys were built, programmed, and appended to the sites Musculoskeletal Questionnaire within clinician group's electronic health record (EHR) systems. As a result, our implementation and testing mirror the workflow of many orthopedic practices utilizing PROMs, contributed to the low missing data.
      Reviewer 11 inquired about how audits can be conducted to ensure a representative sample of patients is used for the clinician group level analysis. The CGA PROM is designed to be used with all eligible patients, not a sample. The pre-surgical and post-surgical PROMs can both be scored independently on the individual level, and both have clinical utility for individual patient counseling. The CGA surveys are not designed as separate PROM that practices need to administer in order to complete a PRO-PM, but as meaningful additions to clinical practice with all patients.
      Reviewer 2 was concerned that when patients select the response option "does not apply to me," that item may be treated as missing or dropped from analysis, resulting in smaller sample sizes. However, the PROM scoring does not drop these cases but has provisions for generating a score from them. Please see section sp.22 of our application, which details the PROM/PRO-PM scoring procedure, including for "does not apply to me" responses.
- **Issue 5: Unidimensionality is not established -** In question 17, reviewer 5 had concerns about our conclusion of unidimensionality for the PROM scores, given the factorial complexity of the EFA results.
    - **Developer Response 5:** Thank you for taking the time to review our submission and providing feedback about unidemensionality. Reviewer 5 was concerned about our conclusion of unidimensionality for the PROM scores, given the factorial complexity of the EFA results. Our reasons for calling it unidimensional were listed in detail in our application.
      We would like to refer back to our response to issue 9 under reliability. The factor structure of the PROM instruments only relates to their use as a psychometric scale for patient assessment. But in the context of the CGA PRO-PM, the 8 items on the surveys are scored entirely differently, starting with a comparison of the pre- and post-responses and ending in a binary pass/fail for each patient. In this measurement scheme, the factor structure underlying the 8 items is not relevant.
- **Issue 6: Was the method described and appropriate for assessing the accuracy of ALL critical data elements? (Question 13; checkboxes)**

- **Developer Response 6:** Thank you for taking the time to review our submission and providing feedback on this specific issue. We noticed that 'Yes', 'No' and 'Not applicable' were selected for question 13. While we knowledge the various opinions of the reviewers, we think our methodology was appropriate for assessing the validity of ALL critical data elements**.** Please see our response in issue 3 of the validity section which addresses the accuracy of ALL critical data elements.
  As we stated in the opening comments and other sections, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM.
  Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality, we conducted the appropriate testing for the PRO-PM.

- **Issue 7: Method of establishing validity at the accountable-entity level (Question 14; checkboxes)**
  - **Developer Response 7:** Thank you for taking the time to review our submission and providing feedback on this specific issue. We noticed that 'Face validity', 'Empirical validity' and 'N/A' were selected for question 14. While we knowledge the various opinions of the reviewers, we think our methodology was appropriate for establishing validity on the accountable-entity level. Please see our response in issue 3 of the validity section which addresses the issue of establishing validity at the accountable-entity level.

    As we stated in the opening comments, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality it had been conducted and completed on the appropriate level.

- **Issue 8: Was the method described and appropriate for assessing conceptually and theoretically sound hypothesized relationships? (Question 15; checkboxes)**
  - **Developer Response 8:** Thank you for taking the time to review our submission and providing feedback on this specific issue. We noticed that 'Yes', 'No' and 'Not applicable' were selected for question 15.

    As we stated in the opening comments, we noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. Please see our response in issue 3 of the validity section which addresses this validity concerns.
    Consistently, we do have some worries about the reviewer's preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes, including validity testing.

- **Issue 9: Assess the method(s) for establishing validity –** In question 16, there were reviewers 2 and 11 who had concerns about the sample size of the clinician groups.

- **Developer Response 9:** Thank you for taking the time to review our submission and providing feedback on this specific issue**.** We noted other concerns related to sample size are also raised in other questions, therefore the response regarding the sample size issues noted by reviewers 2 and 11 are addressed in Issue 1 in the validity section.

- **Issue 10: Assess the method(s) for establishing validity –** In question 16, reviewer 4 acknowledged the TEP, public comment, and comparison of PRO and satisfaction, but questioned if there is a justification for the 75% cutoff, considering we piloted in a population with >50% college grads?
  - **Developer Response 10:** Thank you for taking the time to review our submission and providing feedback on the specific issue of the 75% cutoff for our measure. The measure counts all patients who score 75% or higher on expectations met or exceeded, as they are included in the numerator and indicate having higher care goal achievement.
    The 75% threshold is based on environmental scans of relevant PROMs and PRO-PM scoring and scoring types, input from patient and providers via interviews, conversations with measure developers, a statistician, and a psychometrician, and our PROMs and PRO-PM data. The data for the clinician-groups combined showed an average of 48.5% of patients meeting the 75% threshold, which demonstrated that the threshold was not only obtainable but also allowed for improvement.
    The 75% threshold is also conceptually analogous to the Patient Acceptable Symptomatic State (PASS) score, whereby Tubach (2005) determined 75 is the centile of a final score in patients who consider their health state to be satisfactory, i.e., the highest level of symptoms beyond which patients consider themselves well. It provides clinically meaningful information to interpret results from scales or questionnaires.
    The PASS score also looked at educational status as a possible influence on overall health status and was unable to show any impact on PASS status, therefore we are comfortable with the varying educational status of our population.
    The 75% threshold was vetted and endorsed by the technical expert panel (TEP) members, the project's Steering Committee which include orthopedic surgeons familiar with orthopedic PROMs, and other key stakeholders.
- **Issue 11: Assess the method(s) for establishing validity –** In question 16, reviewer 9 had concerns about the face validity via TEP and public comment and level of expertise.
  - **Developer Response 11:** Thank you for taking the time to review our submission and providing feedback on the specific issue of face validity. Our measure was submitted for a public commenting period to a group of experts (other than those in the TEP) to rate the face validity of the measure specifications. Out of the 10 participants, 10 (100%) agreed at the highest level that the scores from the measure as specified would provide an accurate reflection of quality and value to assess care goal achievement before and after total hip arthroplasty (THA) and/or total knee arthroplasty (TKA).
    The 10 participants from the public comment were experts from a wide variety of backgrounds - measure developers (n=3), medical associations (n=1), healthcare leadership (n=2), surgeons (n=1), EHR vendors (n=2), and healthcare policy makers (n=1). There were multiple orthopedic surgeons, both as public commenters and TEP members who were included in the testing.
- **Issue 12: Assess the method(s) for establishing validity –** In question 16, reviewer 3 had concerns about the data element validity testing method as it related to the PROMs.
  - **Developer Response 12:** Thank you for taking the time to review our submission and providing feedback on this specific issue. As we stated in the opening comments, we

noticed that some of the reviewers' feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM.

Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality it had been conducted and completed on the appropriate level.

- **Issue 13: Please describe any concerns you have with measure exclusions –** In question 18, reviewer 2 questioned how many patients were excluded because they were unable to self-report.
    - **Developer Response 13:** Thank you for taking the time to review our submission and raise this specific issue. Overall, almost all of the reviewers had no concerns about overall measure exclusion, however, reviewer 2 did ask a question for more clarity about how many patients were excluded because they were unable to self-report. We utilized CPT/ICD-10 codes to calculate our exclusions, with the numbers being small - only the size of 3% of the full THA sample and 7% of the TKA sample.

      Due to the subjective nature of the care goal achievement concept and measure, it is recommended that the patient undergoing the THA or TKA is the sole responder to the CGA surveys. Therefore, we do not recommend proxy responses be allowed on behalf of the patient. The decision about patients being the sole responder, and the recommendation against proxy responses, were based on input from key stakeholders, including patients and providers. Importantly, this topic **was** discussed and approved by the TEP members.
- **Issue 14: Risk Adjustment (Questions 19b,c, and d; checkboxes)**
    - **Developer Response 14:** Thank you for taking the time to review our submission and raise this specific issue of risk adjustment. The various ratings that were selected included 'Moderate,' 'Low' and 'Insufficient.' As we stated in the opening comments, we noticed that some of the reviewers feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM.

      Consequently, we have concerns that it would appear that we did not conduct the appropriate testing when in reality it had been conducted and completed on the appropriate level.
- **Issue 15: Please describe any concerns you have regarding the ability to identify meaningful differences in performance -** In question 20, several reviewers (2,4,5,7,9,10,11) had concerns about the sample size in relationship to the meaningful differences in performance.
    - **Developer Response 15:** Thank you for taking the time to review our submission and raise this specific issue. We believe these concerns are related to sample size, which was raised in other questions, therefore the response regarding this concern is addressed under Issue 1 in the Validity section.
- **Issue 16: Please describe any concerns you have regarding comparability of results if multiple data sources or methods are specified –** In question 21, reviewer 2 and 5 raised questions about the process for linking claims/ EHR data for risk models with survey data not being well described. Reviewer 5 felt that the data source(s) are used for the risk factors was not

mentioned and that validity testing was conducted as to these data.

- **Developer Response 16:** Thank you for taking the time to review our submission and raise these issues about data sources and risk factors. The PROMs responses, which include other measures like the HOOS-PS, KOOS-PS, and the PROMIS-10 are collected via a patient-user interface and the data is stored in an enterprise data warehouse (EDW). Patient demographic data such as age, gender, BMI, surgery date and physician appointments are also stored in the data warehouse.
  The risk factors used for risk adjustment - age, gender, and BMI - are standard elements captured for surgical patients. The additional data elements used for reporting are part of the process of care and billing, (i.e., age, gender, BMI, diagnosis, procedure type, etc.), and should be available for all participating clinician-groups.
  Data related to the THA or TKA surgery, like billing data, including the ICD-10 codes and CPT codes are also stored in the EDW. Furthermore, patient risk-adjustment or demographic variables are also stored in the EDW**.**

- **Issue 17: Overall Rating of Validity (Question 25; checkboxes)**
  - **Developer Response 17:** Thank you for taking the time to review our submission and raise these issues about validity testing. It looks like reviewers had different opinions regarding the overall rather of validity. The various ratings included Moderate, Low, and Insufficient.
    While we acknowledge and appreciate the various opinions of the reviewers, we think that we used the appropriate validity testing methods and provided clarifications to their comments and concerns. Please review our responses to the various issues raised by the reviewers in the reliability sections.
    As we stated in the opening comments and other sections, we noticed that some of the reviewers feedback was specific to the development of the PROMs, some was specific to the PRO-PM, and some was specific for both the PROMs and the PRO-PM. Based on our review, it appears that some of the preliminary feedback and decisions were made based on the review of only the PROMs testing methodology and outcomes and not the PRO-PM. As a result, we do have some worries about the reviewer's preliminary feedback about the appropriateness of our PRO-PM testing methodology and outcomes and consequently, the overall rating of the reliability.
- **Issue 18: Rationale for Rating of Overall Rating of Validity (Question 26)**
  - **Developer Response 18:** Thank you for taking the time to review our submission and raise these issues about validity testing. It looks like reviewers had different opinions regarding the overall rating of reliability. The various ratings that were selected included Moderate, Low and Insufficient.
    While we acknowledge and appreciate the various opinions of the reviewers, we think that we used the appropriate validity testing methods and provided clarifications to their comments and concerns. Please review our responses to the various issues raised by the reviewers in the reliability sections.

Other General Comments

The measure development team would like to thank the Scientific Methods Panel and its reviewers for reviewing our measure and providing detailed feedback. We appreciate the consideration of our measure.

Measure Number: 3639

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

Measure Developer/Steward: Yale CORE/Centers for Medicare & Medicaid Services

Reliability

- **Issue 1:** A Reviewer noted that specific protocols (e.g. follow-up for non-responders, allowing for mixed modes within practices, etc) would be helpful since the data collection allows for variable modes of administration. They state that the specifications require recording of who completed the survey (survey/proxy). They raise whether the measure requires collection of which language the patient responded in to assess potential proxy or language/literacy biases.
    - **Developer Response 1:**
        - Regarding data collection, currently, this measure allows clinicians and clinician groups to collect data using a range of methods, including paper and electronic formats. This measure utilized the patient-reported outcome measure (PROM) and risk variable data voluntarily submitted as part of the voluntary PRO data collection within the Comprehensive Care for Joint Replacement (CJR) Model. The CJR Final Rule did not define specific protocols regarding data collection; therefore, hospitals could choose the best approach for them and determine protocols for non-responders or implementation of mixed modes (paper, telephone, electronic). During conversations with providers and patients about PRO-PM implementation, stakeholders noted a preference to allow for multiple collection modes to adapt to needs of patient populations (e.g., a patient without access to a computer would not be able to fill out a survey only sent via email but may be able to respond if the survey was collected in person).
        - Regarding collection of who completes the survey, it is correct that who completed the survey (patient or proxy) was an element collected as part of the voluntary PRO in the CJR Model and available in our testing data. The option of completing a survey via a proxy (e.g., a caregiver) was provided in CJR to allow for flexibility for patients and help maximize responses.
        - Regarding language, the HOOS, JR and the KOOS, JR PROMs have only been available in English language, to date. The CJR Final Rule did not include collection of a data element regarding primary language of the patient or respondent. Since the commentor asked about literacy, it is important to clarify that data on a patients' health literacy (using the SILS2) was collected, and this variable is included in the risk-adjustment model.
        - Regarding the comment about missing modes of data collection, the modes of data collection variable was not a required data element defined in the CJR Rule. Table 1 below shows information about the collection mode in CJR (this data was based on Performance Year [PY] 4).

Table 1. Summary Of Data Collection Mode (CJR PY4)

Collection Mode	Percent
Paper	49.7

Collection Mode	Percent
Telephone (active interactive voice response)	7.1
Electronic (web-base, EHR, etc.)	26.7
Missing	16.5

- **Issue 2:** A Reviewer noted that the exclusions listed in Sp.14 are not defined as to the data sources (eg inpatient claims) and the specifications within the data source (specific International Classification of Diseases, 10th Revision, ICD-10 codes).
    - **Developer Response 2:** The measure denominator uses the following inclusion criteria:
        - Enrolled in Medicare Fee-for Service (FFS) Part A and Part B for the 12 months prior to the date of the index admission and enrolled in Part A during the index admission. This criterion is defined using beneficiary enrollment data.
        - Aged 65 or older. This criterion is defined using beneficiary enrollment data.
        - Discharged alive from non-federal short-term acute care hospital. Discharge status and qualifying hospitals are defined using administrative claims data.
        - Having a qualifying elective primary THA/TKA procedure during the index admission. These criteria are defined using administrative claims data. Elective primary THA/TKA procedures are defined as those THA/TKA procedures without any of the following:
            - Fracture of the pelvis or lower limbs coded in the principal or secondary discharge diagnosis fields on the index admission claim
            - A concurrent partial hip or knee arthroplasty procedure
            - A concurrent revision, resurfacing, or implanted device/prosthesis removal procedure
            - Mechanical complication coded in the principal discharge diagnosis field on the index admission claim
            - Malignant neoplasm of the pelvis, sacrum, coccyx, lower limbs, or bone/bone marrow or a disseminated malignant neoplasm coded in the principal discharge diagnosis field on the index admission claim
            - Transfer from another acute care facility for the THA/TKA
    - We updated the data dictionary to include the specific codes and code placement of the ICD-10 diagnosis codes, ICD-10 procedure codes, and/or present on admission information used to define a qualifying elective, primary THA/TKA procedures (Tab Elective Primary Procedures).
- **Issue 3:** Reviewers selected the following data sources for the measure: claims, abstracted from electronic health records, instrument-based data, and enrollment data.
    - **Developer Response 3:** To clarify, the measure uses claims (for confirmation of primary elective THA/TKA procedures, for identification of some risk factors), enrollment data (used to assess Medicare Fee-for-Service (FFS) enrollment and race), and instrument-based PRO data. Although the CJR data allowed for electronic capture of the surveys, the electronic capture of surveys represented a proportion of the overall data (26.7%).
- **Issue 4:** A reviewer noted that they accept the summary of the validity information presented on the individual instruments.

  - **Developer Response 4:** We provide the reliability and validity testing performed during the development of the HOOS, JR and KOOS, JR in the testing form.
- **Issue 5:** A reviewer expressed confusion over the method for the ICC analyses for test-retest reliability.
  - **Developer Response 5:** We believe this comment pertains to the measure score reliability. To clarify, we did not perform test-retest reliability, we used signal-to-noise reliability which is defined in question 2a.10.
- **Issue 6:** A reviewer raised a concern that reliability testing was not conduced for all measure data elements; instead, it was performed for certain data elements and for measure score reliability.
  - **Developer Response 6:** Since the measure outcome is defined using the HOOS, JR and KOOS, JR we focused our reliability testing results on those elements. Regarding Medicare Claims data, data element reliability of the codes identified through Medicare claims data, which were used to define the cohort and for risk adjustment, is expected as a result of the routine auditing of these billing data by CMS. CMS has in place several hospital auditing programs used to assess overall claims code accuracy, ensure appropriate billing, and for overpayment recoupment. CMS routinely conducts data analysis to identify potential problem areas and detect fraud and audits important data fields used in our measures, including diagnosis and procedure codes, and other elements that are consequential to payment.
  - Likewise, the risk variables collected with PRO data that are included in the risk model were defined during the hospital-level THA/TKA PRO-PM development, which spanned over several years through multiple, iterative steps that incorporated stakeholder input on feasibility, clinical capture, accuracy, reproducibility, and clinical face validity. These steps included: surveying orthopedic practices regarding the feasibility, uniformity, and reliability of risk variables identified by clinical experts and published literature; attending a consensus summit by orthopedic specialty societies to narrow and prioritize clinical risk variables for prospective collection as part of the CJR model (these recommendations were adopted in toto by CMS); seeking additional clinical and empirical evaluation of CJR data; and receiving TEP approval.
- **Issue 7:** A reviewer asked for details about descriptions on variation in response by variable and the results of the non-response bias approach (including coefficients). The reviewer expressed that race was included in the non-response bias adjustment but had a small coefficient when added to the risk model analysis. The reviewer expressed concern that results within racial groups might influence nonresponse and the sample of black patients among responders may be unrepresentative of all black patients undergoing THA/TKAs. They also noted that in general, non-white patients are underrepresented in the population having a THA/TKA.
  - **Developer Response 7:** Please see Table 2 below which shows the characteristics of procedures with Complete, Incomplete, and no PRO or risk variable data within our testing dataset. To clarify, race was not included in the final risk adjustment model; initial analysis exploring the potential impact of race on the risk model provided important information for measure development decisions, but race is not used in the risk model and instead is included in the propensity score model addressing potential non-response bias. The final decision to include race in the approach to non-response

bias was due to evidence in the literature and in the data that persons of non-white race might be underrepresented in the data**.**

- You are correct that when non-white race was individually evaluated in the risk model, the c-statistic was unchanged (Table 17). Regarding how the population of non-white patients compared to the Medicare population undergoing THA/TKA in general, w**e** examined the elective, primary THA/TKA Medicare population between April 2017-March 2020 and found that 9% of THA/TKA patients were non-White. In our testing dataset, we found slightly lower percentages of non-White (7.6%). Of note, the data available for testing of this measure was from the Comprehensive Care for Joint Replacement (CJR) Model participating hospitals that submitted voluntary patient-reported outcome measure (PROM) data, and may not sufficiently represent the national population. Given the known variation in response rates to PROs due to social risk factors, our statistical approach to potential response bias applies weighting based on important factors such as race and dual eligibility (as well as Agency for Healthcare Research and Quality [AHRQ] socioeconomic [SES] index). 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 8:** A reviewer stated that the signal-to-noise reliability results were moderate to high and requested more information on the characteristics of patients who did not provide PROM data.
  - **Developer Response 8:** Table 2 below include the characteristics of patients with THA/TKA procedures with complete PRO data (these are the patients included in the measure), incomplete data (including submissions with missing data elements and submissions of only preoperative PRO data or only postoperative PRO data), and patients with no PRO data (non-response). For an explanation of the statistical approach to potential non-response bias, please see question 2b.08. Please note that among non-responders, only data from Medicare administrative claims is available, and those risk variables collected with PRO data cannot be reported.

Table *2*. Characteristics of Procedures with Complete, Incomplete, and no PRO or risk variable data

Characteristics	Complete PRO, N (%)	Incomplete PRO, N (%)	Non-response, N (%)
Total Admissions	19,429	17,220	41,012
Age Mean (SD)	73.72 (5.73)	73.74 (5.84)	73.83 (5.86)
Male	7,294 (37.54%)	6,291 (36.53%)	15,343 (37.41%)
BMI Mean (SD)	30.27 (5.96)	29.41 (7.96)	-
Mental Health Score Mean (SD)	50.00 (8.10)	48.04 (7.73)	-
Index admissions with an elective THA procedure	6,971 (35.88%)	5,920 (34.38%)	13,971 (34.07%)
Index admissions with an elective TKA procedure	12,458 (64.12%)	11,300 (65.62%)	27,050 (65.96%)
Number of procedures (two vs. one)	116 (0.60%)	180 (1.05%)	1422 (3.47%)
Race: Unknown	315 (1.62%)	335 (1.95%)	733 (1.79%)
Race: White	17,946 (92.37%)	15,572 (90.43%)	37,128 (90.53%)

Characteristics	Complete PRO, N (%)	Incomplete PRO, N (%)	Non-response, N (%)
Race: Black	681 (3.51%)	772 (4.48%)	1748 (4.26%)
Race: Other	189 (0.97%)	201 (1.17%)	476 (1.16%)
Race: Asian	137 (0.71%)	142 (0.82%)	426 (1.04%)
Race: Hispanic	101 (0.52%)	143 (0.83%)	413 (1.01%)
Race: North American Native	60 (0.31%)	55 (0.32%)	88 (0.21%)
Low SES: lowest quartile of AHRQ SES*	1,833 (9.43%)	1,785 (10.37%)	4,138 (10.09%)
Dual Eligibility**	539 (2.77%)	655 (3.80%)	1,855 (4.52%)
Severe infection; other infectious diseases (CC 1, 3-7)	3,409 (17.55%)	3,172 (18.42%)	7,328 (17.87%)
Diabetes mellitus (DM) or DM complications (CC 17-19, 122-123)	5,018 (25.83%)	4,596 (26.69%)	11,050 (26.94%)
Liver disease (CC 27-31)	813 (4.18%)	786 (4.56%)	1,778 (4.34%)
Rheumatoid Arthritis and Inflammatory Connective Tissue Disease	2,083 (10.72%)	1,828 (10.62%)	4,322 (10.54%)
Depression	3,012 (15.50%)	2,695 (15.65%)	6,511 (15.88%)
Other Psychiatric Disorders	3,099 (15.95%)	2,842 (16.50%)	6,685 (16.30%)
Coronary atherosclerosis or angina (CC 88-89)	4,750 (24.45%)	4,428 (25.71%)	10,268 (25.04%)
Vascular or circulatory disease (CC 106-109)	3,727 (19.18%)	3,446 (20.01%)	8,194 (19.98%)
Renal failure (CC 135-140)	2,753 (14.17%)	2,496 (14.49%)	6,121 (14.92%)
Healthy Literacy: Not at all	3,282 (16.89%)	2,690 (15.62%)	-
Healthy Literacy: A little bit	1,502 (7.73%)	1,535 (8.91%)	-
Healthy Literacy: Somewhat	2,124 (10.93%)	2,262 (13.14%)	-
Healthy Literacy: Quite a bit	3,489 (17.96%)	3,328 (19.33%)	-
Healthy Literacy: Extremely	9,032 (46.49%)	7,118 (41.34%)	-
Healthy Literacy: Missing Literacy	0 (0.00%)	287 (1.67%)	-
Other Joint Pain: None	6,694 (34.45%)	5,221 (30.32%)	-
Other Joint Pain: Mild	4,768 (24.54%)	4,069 (23.63%)	-
Other Joint Pain: Moderate	4,897 (25.20%)	4,664 (27.08%)	-
Other Joint Pain: Severe	2,516 (12.95%)	2,437 (14.15%)	-
Other Joint Pain: Extreme	554 (2.85%)	639 (3.71%)	-
Other Joint Pain: Missing	0 (0.00%)	190 (1.10%)	-

Characteristics	Complete PRO, N (%)	Incomplete PRO, N (%)	Non-response, N (%)
Back Pain: None	7,328 (37.72%)	6,455 (37.49%)	-
Back Pain: Very Mild	4,884 (25.14%)	4,109 (23.86%)	-
Back Pain: Moderate	4,988 (25.67%)	4,289 (24.91%)	-
Back Pain: Fairly Severe	1,601 (8.24%)	1,519 (8.82%)	-
Back Pain: Very or Worst Severe	628 (3.23%)	639 (3.71%)	-
Back Pain: Missing	0 (0.00%)	209 (1.21%)	-
Use of Chronic (>= 90 days) Narcotics***	3,390 (17.45%)	3,180 (18.47%)	-

*Note: Missing AHRQ SES Index information among the Complete PRO population was (0.21%), 0.25% among the incomplete PRO population, and 0.19% for the non-response population.
**Note: Missing Dual Eligibility information among the non-response population was 0.01%.
***Note: Missing use of Chronic Narcotics 0.51% in the incomplete PRO population.

Validity

- **Issue 1:** A reviewer noted that entity-level validity was not conducted for the measure.
    - **Developer Response 1:** Thank you for your question.
- **Issue 2:** A reviewer noted that not all data elements included in the measure had validity testing and requested more details regarding the face validity process.
    - **Developer Response 2:** As noted above, since the measure outcome is defined using the HOOS, JR and KOOS, JR we focused our validity. testing results on those elements. The TEP consisted of 20 total members, five of which were patients, 11 clinicians, including 6 orthopedic surgeons, and 4 experts in performance measurement while the Patient Working Group consisted of six patients who have undergone hip and/or knee procedures. We have engaged both groups throughout measure development and testing.
    - Regarding recruitment, TEP members were selected through a publicly posted call for TEP on the CMS website and patients were recruited through partnerships with Rainmakers.
    - Feedback was obtained via teleconference calls. Patients engaging in this work were provided with preparation calls that reviewed the meeting materials ahead of the meeting date and debrief calls that allowed them to share any thoughts after the scheduled meeting. All meeting materials were sent in advance to allow individuals time to review the performance results and data.
    - To date, the TEP has provided input on and supported the measure concept, clinician and clinician group attribution of THA/TKA procedures, and risk model approach and results. In addition, we reviewed the approach to social risk factor analyses and results, approach to response bias and results, the final measure scores and reliability and validity testing. We also reviewed future measure specifications updates, to expand the measure cohort and extend the postoperative PROM data collection window.
    - The Patient Working Group provided input on and supported the measure concept, measure use, and approach to analyzing social risk and nonresponse bias. We also discussed future measure specifications updates.

- We asked the TEP to respond to the face validity questions via a poll during a TEP Meeting or for those unable to join the call or complete the poll, we followed up via phone or email. We asked the patient working group members to respond to the face validity statements via survey. For both the TEP and patient working group, we explained that the face validity statements are routinely collected during the measure development process and would be included in publicly available documents.

- **Issue 3:** A reviewer expressed concern regarding the PRO submission rates and ceiling effects of the HOOS, JR.
  - **Developer Response 3:** The true "response" rate for our study is difficult to calculate because it is unknown to whether 100% of eligible patients in our dataset were asked to provide PRO data. However, we do have the true denominator of eligible cases, based upon claims data. In the absence of a true "response" rate, we have calculated an estimated response rate as the percentage of all elective primary THA/TKA procedures meeting cohort criteria performed during the measurement period by all the clinicians and clinician groups in the dataset for which complete and matched preoperative and postoperative PRO and risk variable data were submitted. With this operational definition, the mean response rate across clinicians was 32.23% (SD 24.55%) and 31.85% (SD 24.20%) across clinician groups. Among clinicians with ≥25 elective primary THA/TKA patients with complete PRO data during the measurement period, the mean response rate was 42.09% (SD 16.98%); among clinician groups with ≥25 elective primary THA/TKA patients with complete PRO data during the measurement period, the mean response rate was 36.65% (SD 18.38%) (see Tables 9 and 10 in the Testing Form).
  - Please note that response rates may have been impacted by hospital submission thresholds set by CJR. The CJR model within which these PRO data were collected, required that hospitals submitting the data meet either a minimum percentage or an absolute minimum number of PRO cases to qualify for the quality point incentive; the thresholds in CJR performance years one, two, three, and four were 50% of or 50 eligible cases; 60% of or 75 eligible cases; 70% or 100 eligible cases; and ≥ 80% or ≥ 200 eligible procedures, respectively. To address potential response bias, we used stabilized inverse probability weighting, created with a multinomial logistic regression to calculate stabilized inverse probability weights.
  - Regarding the ceiling effects, this measure uses the same substantial clinical benefit (SCB) threshold developed for the hospital-level measure, which was reviewed and recommended for endorsement by the NQF Surgery Standing Committee in 2020. SCB improvement is defined as follows:
    - SCB thresholds were defined using published literature (Lyman and Lee, 2018) and vetted by the hospital-level THA/TKA PRO-PM development Patient Working Group, Technical Expert Panel (TEP), Technical Advisory Group, and Orthopedic Clinical Expert.
    - An improvement threshold approach 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.
- **Issue 4:** A reviewer expressed concern about how well the nonresponse bias adjustment is performing.
  - **Developer Response 4:** The comparison of clinician-level and clinician group-level RSIRs for a risk-adjusted model of SCB improvement with stabilized IPW and without stabilized

IPW suggests that the results are not sensitive to our weighting adjustment. We will want to continue to evaluate this non-response bias approach. Due to the high proportion of non-responders, we consider it important to account for the differences in characteristics of responders and non-responders found in the literature and empirically in our data. We expect non-response bias will be a factor for the this measure due to associations with non-response including SES and health status. We, therefore, retained response bias adjustment for the measure results.

- **Issue 5:** A reviewer noted concern regarding the proportion of patients excluded from the cohort based on missing data or not being attributed to a clinician.
    - **Developer Response 5:** We recommend CMS continue to evaluate response rates and rates of attribution to clinicians in the future. Please note that a majority of patients excluded from the cohort were due to non-response. As noted in Response #3 above, , the mean response rate across clinicians was 32.23% (SD 24.55%) and 31.85% (SD 24.20%) across clinician groups. Among clinicians with ≥25 elective primary THA/TKA patients with complete PRO data during the measurement period, the mean response rate was 42.09% (SD 16.98%); among clinician groups with ≥25 elective primary THA/TKA patients with complete PRO data during the measurement period, the mean response rate was 36.65% (SD 18.38%) (see Tables 9 and 10 in the Testing Form). The voluntary nature of the PRO data collection in CJR makes high response rates challenging. The measure implementation will have to carefully consider approaches to increased response rates.

Other General Comments

- **Issue 1:** Reviewers checked off both yes and no for social risk adjustment.
    - **Developer Response 1:** To clarify, based on the results of the social risk factor testing, we did not include additional social risk factors beyond health literacy. As noted above, we do include health literacy in the final risk model, based upon strong patient and technical expert input. In our dataset, neither dual eligibility, AHRQ SES index lowest quartile, nor non-white race was not statistically significantly associated with the outcome, and inclusion of these variables in the risk model did not appear to impact RSIRs. Additional analysis of clinician and clinician group proportion of dual eligible, AHRQ SES index lowest quartile, and non-white race patients by clinicians and clinician groups indicate that those with the lowest proportion of dual eligible, AHRQ SES lowest quartile, and non-white race patients and those clinicians and clinician groups with the highest proportion of these patients have similar RSIR distributions. These data do not provide evidence of significant differences in RSIRs due to the proportion of a hospital's patients with dual eligibility, AHRQ SES lowest quartile, and non-white race. The lack of association and effect of these factors may be due to lower case selection in these groups for these elective primary procedures. We will continue to monitor the impact of social risk factors.
- **Issue 2:** A reviewer noted concern regarding the adequacy of the c-statistic for the risk model and the measure's predictive ability.
    - **Developer Response 2:** We examined the C-statistic for the models per year of data, as well as for both years, and noted differences in the statistic per year: the first full year of PRO data was 0.68, using only the second full year of data revealed a lower C-statistic

(.059). Since the risk model is performed at the patient-level, we do not anticipate the clinician attribution would impact the results. Investigation of the patient characteristics between those undergoing a THA/TKA in the first year (July 1, 2016 – June 30, 2017) and those undergoing a THA/TKA in the second year (July 1, 2017 – June 30, 2018) did reveal a strong association between improvement and health literacy in the first year of data that does not persist in the second year of data. We recommend continued assessment of the risk model in a larger dataset in the future. With regard to predictive ability, our testing results indicate good predictive ability, and we recommend CMS continue to assess this in the future.

- **Issue 3:** A reviewer asked for an analysis which identifies the percent of clinicians and clinician groups with statistically higher and lower rates.
    - **Developer Response 3:** Please see the tables 11 and 12 that provide distribution of RSIRs among clinicians and clinician groups.

Measure Number: 3667

Measure Title: Days at Home for Patients with Complex, Chronic Conditions

Measure Developer/Steward: Yale CORE/Centers for Medicare & Medicaid Services

Reliability

- **Issue 1:** Variation by ACO size not shown.
    - **Developer Response 1:**
        - Due to limited data availability, CORE was not able to test in additional years during the development process.
        - Below (in Table 1), we show the mean days at home by quartiles of beneficiary volume; we found that lower-volume ACOs tend to perform slightly better on the measure of days at home, but not meaningfully nor significantly so. Moreover, the variation within each quartile is much greater than any variation between quartiles.

Table 3. Distribution of ACO Number of Days at Home by Quartile of Beneficiary Volume

Quartile by Volume	Mean Days at Home	Standard Deviation
Q1 (56 ≤ n ≤ 792)	331.04	2.68
Q2 (793 ≤ n ≤ 1,191)	330.40	3.24
Q3 (1,194 ≤ n ≤ 2,255)	329.84	5.48
Q4 (2,263 ≤ n ≤ 13,426)	330.22	2.65

- **Issue 2:** Specific form of ICC not specified
    - **Developer Response 2:**
        - CORE used **ICC (2,1)** defined by Shrout & Fleiss (1979):
            - Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Validity

- **Issue 1:** Method for combing three risk adjusted models

- **Developer Response 1:**
  - There are three risk-adjusted models used in the Days at Home measure: Days in Care, Mortality, and Nursing Home admission. The purpose of the latter two models is to avoid potential unintended consequences, such as incentivizing ACOs to withhold appropriate care (which may lead to higher mortality) or to actively transition beneficiaries to non-skilled nursing facilities in an attempt to improve measure performance without first attempting home- and community-based solutions. Briefly, the Days in Care model is used to construct an initial measure for each ACO, and this value is then updated using risk standardized mortality and nursing home transition ratios. Though the use of three models is not typical, we arrived at this as approach in part because it is less complex than trying to account for both survival time and transition to nursing home in a single model for Days in Care, and it was endorsed by both our Technical Expert Panel (TEP) and CMS.
  - Conceptually, consider an ACO that has a high unadjusted mortality rate and a challenging case mix. Penalizing such an ACO for a high rate of mortality may be unfair given that their patient population is more ill in the aggregate than other ACOs. Another ACO could have a low unadjusted rate of nursing home use because it serves a younger population of beneficiaries. Using risk adjustment in each model helps ameliorate these fairness concerns and level the playing field.
  - Specifically, we first use the Days in Care model to estimate for each patient their "excess days in care" (EDIC), representing the number of days they spend in acute care over similar patients at an average ACO. Then, we use the Mortality and Nursing Home models to estimate for each patient their standardized mortality ratio (SMR) and standardized nursing home ratio (SNHR), which are the patients' excess risks of death and transition to long-term nursing home care, respectively, that can be attributed to an ACO. We then standardize SNHR (to rSNHR) and apply a formula to transform these into a number of days that are either added or subtracted from the EDIC, effectively transforming the information we have about mortality and nursing home transitions at that ACO into some number of "extra" Excess Days in Care.
  - For example, if rSNHR and SMR are both higher than expected, this is a negative result, therefore additional days in care are added onto the EDIC model output; this equates to lower Adjusted Days at Home, which is the score that is reported. The EDIC remains the central focus of the measure, and indeed ACOs with average EDIC (that is, equal to zero) are not affected by the SMR and SNHR adjustments, but ACOs with better or worse EDIC will see their scores modified somewhat if they have better or worse nursing home transitions or mortality. Importantly, this approach was endorsed by the Technical Expert Panel as appropriately but not excessively accounting for variation on mortality and nursing home rates across ACOs.
  - Table 2 below illustrates the distribution of "extra" EDIC due to the SMR adjustment and due to the rSNHR adjustment alone.

Table 4. Distribution of change in EDIC due to Mortality & Nursing Home model adjustments

Quantile	Change in EDIC: SMR adjustment	Change in EDIC: rSNHR adjustment
Maximum	+4.13	+12.0
90%	+0.41	+1.34
Q3 (75%)	+0.12	+0.63
Median (50%)	+0.00	-0.03
Q1 (25%)	-0.11	-0.58
10%	-0.45	-0.99
Minimum	-10.2	-5.36

- **Issue 2:** Listed exclusions and exceptions
    - **Developer Response 2:**
        - CORE discussed the outcome definition extensively with the TEP and other experts, and ultimately decided to adopt a broad outcome definition that includes almost any acute or post-acute care use in select settings, with very limited exceptions. Accordingly, the measure does not distinguish between planned/elective vs. unplanned/non-elective inpatient admissions, or between "preventable" vs. "non-preventable" inpatient admissions; all are counted as "days in care." This is based on a strong preference from the TEP, which noted that the distinctions between these categories are not always clear-cut (for example, some unplanned admissions may not be preventable while some planned admissions may not be medically appropriate) or may not always be consistently identifiable in available claims data. Furthermore, most inpatient admissions could be considered disruptive to a patient's daily life and the TEP felt that conceptually few if any should be considered "at home." Several TEP members advocated a "head-in-bed" standard, in which a patient who cannot be in their own bed should be considered "in care" regardless of the circumstance of admission.
        - The measure is intended for use in entities, such as ACOs, that have taken on a commitment to full-person care for their aligned patients. The intent is not to eliminate inpatient care for patients of measured entities, but rather to create an incentive for entities to explore or innovate methods to deliver effective home- and community-based care that produces good outcomes while reducing their patients' needs for disruptive and higher-risk acute care. The inclusion of the mortality adjustment to the score is intended partly as a balance against unintended consequences, to introduce a penalty to providers who reduce their patients' acute care use in irresponsible ways while rewarding those who are still able to deliver high-quality outcomes.
        - Finally, to clarify, the measure does not include any denominator exceptions – all patients meeting the stated inclusion criteria (adults aligned to measured entities at the start of the year, continuously enrolled in Medicare Fee-for-Service [FFS], and with a Hierarchical Condition Category [HCC] score of 2.0 or greater) are included and their patient-days are included in the denominator. There are limited exclusions from the numerator (namely, select obstetric

admissions and care delivered while enrolled in hospice), but these patient days are still counted in the denominator and these patients are eligible to have counted days in care that do not meet those numerator exclusions.

- **Issue 3:** Arbitrary weighting of mortality and nursing home adjustments
    - **Developer Response 3:**
        - The TEP and other stakeholders stated the importance of representing mortality as more severe than nursing home transition. Our decision to weight the adjustment in this way aligns with the view that mortality should have a greater impact on the measure score than nursing home transitions. It is important to note that for many ACOs with near-average Days in Care, the impact of the mortality and nursing home transition adjustments is minimal; these adjustments primarily affect providers with very good or very bad performance on either of those two components.
        - While the weights are in some sense arbitrary, they are constructed so that for any two ACOs with similar excess days in care, the one with lower mortality rate or lower nursing home rate will have a better final score. Though the impact of both adjustments on the score are quite modest, we believe they provide appropriate protection against unintended incentives.
- **Issue 4:** Definition of "complex chronic conditions"
    - **Developer Response 4:** CORE defined a criterion of HCC score of 2.0 or greater to define the cohort of "complex, chronic conditions." This aligns with criteria in two CMS Innovation Center models that plan to use the measure: the Primary Care First model's "seriously ill patient" definition and the Direct Contracting model's "high-needs patient" definition. The HCC score is readily available in claims data, straightforward to implement in calculation logic, and generally reflects patients with some combination of chronic conditions and/or comorbidities that indicate a greater risk for acute care need compared to the average Medicare FFS patient, but does not restrict the cohort to any specific chronic condition or combination thereof.
- **Issue 5:** Terminology and distinction between "ACOs" and "medical groups"
    - **Developer Response 5:** The measure is intended for use in Accountable Care Organizations (ACOs) or similar organizations, such as Direct Contracting Entities (DCEs) or Primary Care First clinician groups that have accepted responsibility for whole-person care of aligned beneficiaries. In this context, we conceive of an ACO as an organization of providers spanning a continuum of care, all of whom are mutually responsible for the outcome of their aligned beneficiaries. The measure was tested specifically in Shared Savings Program ACOs due to data limitations – since the Direct Contracting and Primary Care First are new payment models, performance data from participants are not yet available. CORE intends additional future testing when those data become available to confirm reliability and validity.
- **Issue 6:** Use of statistical model for mortality
    - **Developer Response 6:**
        - The Days in Care model does not directly include days after death as either days in care or days at home, but is designed to account for the missing information

that results from censoring observation of those patients by including an offset of the number of days alive for each patient who died.
        - CORE, along with the TEP and other experts consulted during the development process, did not support neglecting death as an outcome, as it is clearly a bad outcome for the patient (in general, with the exception of patients in hospice) and provides information about the quality of the provider. The TEP in particular strongly preferred a measure that summarizes "days alive and at home" rather than merely "days at home while alive."
        - With this feedback, CORE's goal was to avoid creating a measure that inadvertently rewards providers with better measure scores for reducing patients' days in care in ways that put their patients at greater risk. At the same time, we did not want to treat raw mortality as a negative outcome as that might discourage providers from treating very ill patients who are at a higher initial risk of mortality. Our method, to compute a risk-adjusted measure of mortality risk and use this to adjust the Days at Home score, was intended to balance these concerns.
- **Issue 7:** Social Risk Factors
    - **Developer Response 7:**
        - To clarify, the Days at Home measure does include risk adjustment for dual-eligible status in two of the three models, the Days in Care and Nursing Home Transition component models. The rationale for this is that dual-eligible status is significant in these models, the data are readily available to use, and there are conceptual reason to believe that dual-eligibility status affects these outcomes either directly (in the case of nursing home transitions, as Medicaid pays for long-term nursing home residence while Medicare does not) or indirectly (as a proxy indicator for constructs like financial status or social supports that affect a person's availability to remain safely at home). This adjustment in these models is intended to avoid discouraging providers from taking on these patients who are at greater risk for poor outcomes in those models due to factors that may be out of the provider's control.
        - We did not include dual-eligible status in the Mortality component model, as mortality may be more influenced by clinical risk factors and this is consistent with other CMS mortality measures.
- **Issue 8:** Meaningful Differences
    - **Developer Response 8:**
        - As noted in the measure submission, the ACO-level mean of adjusted days at home (n=610) was 330.38 (standard deviation [SD] 3.72, interquartile range [IQR] 329.12 – 332.11). At the beneficiary level, the mean of adjusted days at home (n=1,154,779) was 330.22 (SD 4.63, IQR 329.33 – 331.58). In general, the beneficiary-level scores are clustered more closely around the mean (resulting in the narrower interquartile range), but have more extreme outlying values (resulting in the greater standard deviation); overall the variation observed at the ACO level is comparable to the variation observed at the beneficiary level.

- At the beneficiary level, the mean unadjusted Days in Care is 12.76 (SD 25.68, IQR 0 – 12) and the mean unadjusted number of days alive is 345.01 (SD 67.44, IQR 365-365 [that is, more than 75% of beneficiaries survive]). At the ACO level, the mean unadjusted Days in Care is 12.72 (SD 3.02, IQR 10.92 – 13.97) and the mean unadjusted number days alive is 345.94 (SD 5.37, IQR 342.55 – 349.82). As should be expected, there is much greater variation in the raw outcomes at the patient level than at the aggregate ACO level, as there is a distribution of outcomes within each ACO. Of note, the ACO-level SD of unadjusted days in care (3.02) is comparable to that of the adjusted days at home (3.72).
- CORE would like to note that, while this is a nominally a measure of "days at home," the measure itself is calculated based on a model of days in care (as this is more practical to model), with each "day in care" corresponding to one "day not at home." In this context, differences in score should be considered in reference to days in care rather than to all days in the year. For example, an ACO's Days at Home score is 3.7 (one SD) above average, that actually indicates patients of that ACO spent 3.7 fewer days in care on average; this represents 29% of the 12.76 days spent in care by an average patient in the cohort.
- In addition, the final "days at home" score is on a per-patient basis. For example, an ACO with a score that is one standard deviation (3.7 days) above average is providing its patients on average 3.7 additional days home each – for an ACO with 1,000 patients, this is a total gain of 3,700 patient-days at home above expectation. This same ACO by definition is saving its patients on average 3.7 days in care each (for a total savings of 3,700 patient-days in care).
- Regarding the possibility of omitted risk factors explaining the variation, this is a concern of all risk adjusted measures. Unmeasured differences in patient mix across providers are always a possibility. However, one goal of validity testing, especially external validity testing, is to address this concern; our results showing that this measure correlates with other measures of quality suggests that the variation reflects more than omitted risk factors. Omitted risk factors would explain the variation in the measure only if all correlated measures varied according to the presence of the same omitted risk factor.
- Finally, CORE would like to note that while it is difficult to quantify what a difference of a few days in this measure indicates in terms of differences in patient function or health-related quality of life, information obtained from the TEP and the literature indicate that spending even a few days fewer in acute settings and a few days more at home is often valuable to patients who feel safer and more comfortable.

Other General Comments

No additional information or considerations at this time.

Subgroup 2

Measure Number: 0689

Measure Title: Percent of Residents Who Lose Too Much Weight (Long-Stay)

Measure Developer/Steward: Centers for Medicare & Medicaid Services

Reliability

- **Issue 1:** Several panelists expressed interest in seeing the distribution of signal-to-noise reliability scores across facilities.
    - **Developer Response 1:** Since experiencing weight-loss is a binary outcome, reliability was estimated using a beta-binomial model. This model assumes the provider QM score for the weight loss measure is a binomial random variable, conditional on the provider's true value that comes from a beta distribution. Data from 2019Q1 through 2019Q4 were used to conduct this analysis by fitting the beta binomial model to the data. Table 1 below contains the distribution of signal-to-noise reliability scores for the weight loss measure. The average reliability score across all providers was 0.76, and the median score was 0.78, which suggests that the measure is moderately reliable in separating provider characteristics from variability within provider.

Table 5 Distribution of Signal-to-Noise Reliability Scores for Percent of Residents Who Lose Too Much Weight, NQF #0689 (2019Q1-2019Q4)

Average Score	Score Percentile	Score Percentile	Score Percentile
*	**25th**	**50th**	**75th**
0.76	0.68	0.78	0.86

***This cell is intentionally left empty.

- **Issue 2:** One panelist inquired whether or not the split-half reliability was adjusted with the Spearman-Brown (SB) Prophesy formula.
    - **Developer Response 2:** The Split-half reliability statistics calculated are Pearson Correlation (r = 0.64), Spearman correlation (ρ = 0.65), and intraclass correlations (ICC (2,1) = 0.64). The split-half reliability analysis was conducted on all facilities with 40 or more residents counted in the measure denominator across eight quarters (2018Q1 - 2019Q4) to ensure at least 20 residents could be used in each randomly selected half of a facility's residents. Since we conducted a quasi-full-length test by doubling the sample size before splitting the samples, we did not find it necessary to adjust the above results with the Spearman-Brown formula. However, with further adjustment using the Spearman-Brown formula described below, split-half reliability would be 0.78.

$$\textit{Split-half reliability using the Spearman-Brown formula} = \frac{2r}{1+r}$$

Validity

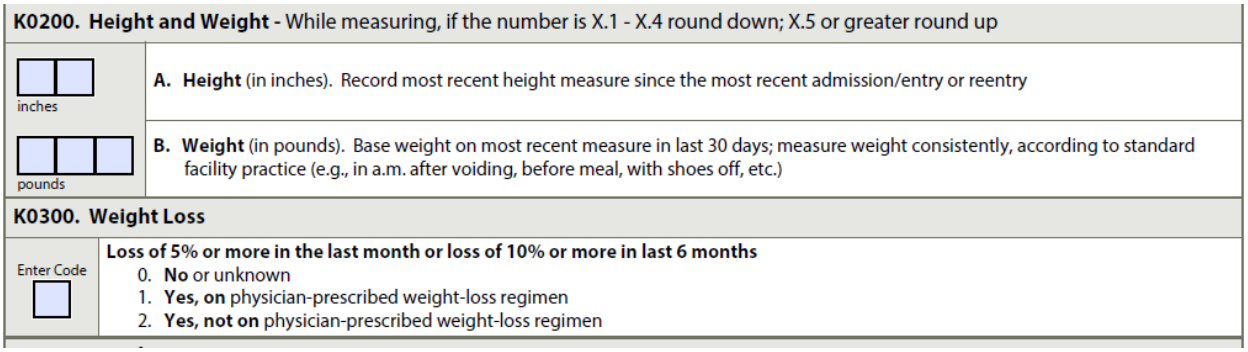
No comments from the developers were submitted.

Other General Comments

- **Issue 1:** One panelist asked for clarification regarding a statement about baseline weight that was made in the measure description.
    - **Developer Response 1:** This measure captures the percent of residents who are not on a physician prescribed weight-loss regimen but experience 5% weight loss within 30 days of the target assessment *or* 10% weight loss within 180 days of the target assessment (K0300 = [2]). The MDS assessment, which is required at least on a quarterly

basis, records resident weight through item K0200B. In addition to K0200B, the MDS manual states that weight should be monitored on a continuing basis. Weight loss should be assessed and care planned at the time of detection. This assessment and care planning should not be delayed until the next MDS assessment. Figure 1 below contains descriptions of the relevant MDS items while Figure 2 describes the MDS guidelines used to assess item K0300 Weight Loss.

Figure 1: MDS Items K0200 and K0300



K0200. Height and Weight - While measuring, if the number is X.1 - X.4 round down; X.5 or greater round up

inches — **A. Height** (in inches). Record most recent height measure since the most recent admission/entry or reentry

pounds — **B. Weight** (in pounds). Base weight on most recent measure in last 30 days; measure weight consistently, according to standard facility practice (e.g., in a.m. after voiding, before meal, with shoes off, etc.)

K0300. Weight Loss

Enter Code

Loss of 5% or more in the last month or loss of 10% or more in last 6 months

0. **No** or unknown
1. **Yes, on** physician-prescribed weight-loss regimen
2. **Yes, not on** physician-prescribed weight-loss regimen

For a New Admission

1. Ask the resident, family, or significant other about weight loss over the past 30 and 180 days.
2. Consult the resident's physician, review transfer documentation, and compare with admission weight.
3. If the admission weight is less than the previous weight, calculate the percentage of weight loss.
4. Complete the same process to determine and calculate weight loss comparing the admission weight to the weight 30 and 180 days ago.

For Subsequent Assessments

1. From the medical record, compare the resident's weight in the current observation period to his or her weight in the observation period 30 days ago.
2. If the current weight is less than the weight in the observation period 30 days ago, calculate the percentage of weight loss.
3. From the medical record, compare the resident's weight in the current observation period to his or her weight in the observation period 180 days ago.
4. If the current weight is less than the weight in the observation period 180 days ago, calculate the percentage of weight loss.

Measure Number: 3633e

Measure Title: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Level)

Measure Developer/Steward: University of California, San Francisco/Alara Imaging, Inc.

Reliability

- **Issue 1: TIME PERIOD OF DATA COLLECTION.** We specified the time period of data collection as "One calendar year, although shorter periods can be used for high-volume entities." One reviewer remarked that we did not define "high-volume."
    - **Developer Response 1:** The measure, as formally specified and submitted, uses one calendar year of data from each accountable entity.

Validity

- **Issue 1: MISSING DATA**. Several reviewers raised concern with the approach of using missing data as a technical exclusion, which could potentially lead to bias. Further, they raise the concern that while missingness was low in our testing data, it may be higher in "real world" implementation.
    - **Developer Response 1:** During testing there was some missing data for 8% of exams and 90% of this was related to missing radiation dose information. **We believe that the issue of missing radiation data is for the most part entirely solvable**. The missing radiation data is not related to an entity's hardware except in very rare situations in which very old machines are used to perform the exam; rather, it is almost entirely a software and data storage issue. The radiation dose data is stored within the Radiation Dose Structured Report (RDSR), a digitized, structured summary of the total radiation output associated with the performance of the CT exam. The RDSR is produced with every CT scan and CMS incentivizes the creation of the RDSR by paying a lower reimbursement for CT scans that do not produce an RDSR. The issue that can arise is that some entities may not save and store the RDSR. There is a widespread campaign organized by the American College of Radiology to encourage entities to save and store RDSR information. Sites that do not currently save the RDSR in their radiology electronic systems will need to invest time and resources in modifying their systems to be able to do so. We calculated the amount of time this requires as part of the testing and it was quite modest, as we will describe in the Feasibility section. Although sites may require vendor support, this work is not excessively burdensome. One of our testing sites went from saving 0% to 96% of their machines' RDSRs in a week's time with remote support from Siemens. Another site with mostly General Electric CT machines increased saving from 10% to 65% within a month, adjusting one machine at a time.
    - As described in section 2b.10, **the measure steward will closely monitor missingness at the accountable entity level and report these numbers to the entities, which will be expected to fix the issue within a reasonable period of time**. If missingness doesn't resolve to near-zero by the time of NQF Maintenance, we will consider revising the measure to establish a missing data threshold beyond which exams with missing data will be treated as a failure.

Other General Comments

- **Issue 1: SOFTWARE COST.** Two reviewers inquired about the cost of the proprietary software that is required to process primary data elements from electronic systems to generate variables required by the eCQM: CT category, size-adjusted radiation dose, and global noise. The reviewers voiced concern about imposing the cost of this software package on accountable entities.

- **Developer Response 1:** Clinicians, clinician groups, and hospitals will be able to report on the measures for free through an open web interface.

Measure Number: 3662e

Measure Title: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Group Level)

Measure Developer/Steward: University of California, San Francisco/Alara Imaging, Inc.

Reliability

- **Issue 1: TIME PERIOD OF DATA COLLECTION.** We specified the time period of data collection as "One calendar year, although shorter periods can be used for high-volume entities." One reviewer remarked that we did not define "high-volume."
  - **Developer Response 1:** The measure, as formally specified and submitted, uses one calendar year of data from each accountable entity.

Validity

- **Issue 1: MISSING DATA.** Several reviewers raised concern with the approach of using missing data as a technical exclusion, which could potentially lead to bias. Further, they raise the concern that while missingness was low in our testing data, it may be higher in "real world" implementation.
  - **Developer Response 1:** During testing there was some missing data for 8% of exams and 90% of this was related to missing radiation dose information. **We believe that the issue of missing radiation data is for the most part entirely solvable**. The missing radiation data is not related to an entity's hardware except in very rare situations in which very old machines are used to perform the exam; rather, it is almost entirely a software and data storage issue. The radiation dose data is stored within the Radiation Dose Structured Report (RDSR), a digitized, structured summary of the total radiation output associated with the performance of the CT exam. The RDSR is produced with every CT scan and CMS incentivizes the creation of the RDSR by paying a lower reimbursement for CT scans that do not produce an RDSR. The issue that can arise is that some entities may not save and store the RDSR. There is a widespread campaign organized by the American College of Radiology to encourage entities to save and store RDSR information. Sites that do not currently save the RDSR in their radiology electronic systems will need to invest time and resources in modifying their systems to be able to do so. We calculated the amount of time this requires as part of the testing and it was quite modest, as we will describe in the Feasibility section. Although sites may require vendor support, this work is not excessively burdensome. One of our testing sites went from saving 0% to 96% of their machines' RDSRs in a week's time with remote support from Siemens. Another site with mostly General Electric CT machines increased saving from 10% to 65% within a month, adjusting one machine at a time.
  - As described in section 2b.10, **the measure steward will closely monitor missingness at the accountable entity level and report these numbers to the entities, which will be expected to fix the issue within a reasonable period of time**. If missingness doesn't resolve to near-zero by the time of NQF Maintenance, we will consider revising the

measure to establish a missing data threshold beyond which exams with missing data will be treated as a failure.

Other General Comments

- **Issue 1: SOFTWARE COST.** Two reviewers inquired about the cost of the proprietary software that is required to process primary data elements from electronic systems to generate variables required by the eCQM: CT category, size-adjusted radiation dose, and global noise. The reviewers voiced concern about imposing the cost of this software package on accountable entities.
    - **Developer Response 1:** Clinicians, clinician groups, and hospitals will be able to report on the measures for free through an open web interface.

Measure Number: 3663e

Measure Title: Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level)

Measure Developer/Steward: University of California, San Francisco/Alara Imaging

Reliability

- **Issue 1: TIME PERIOD OF DATA COLLECTION.** We specified the time period of data collection as "One calendar year, although shorter periods can be used for high-volume entities." One reviewer remarked that we did not define "high-volume."
    - **Developer Response 1:** The measure, as formally specified and submitted, uses one calendar year of data from each accountable entity.

Validity

- **Issue 1: MISSING DATA.** Several reviewers raised concern with the approach of using missing data as a technical exclusion, which could potentially lead to bias. Further, they raise the concern that while missingness was low in our testing data, it may be higher in "real world" implementation.
    - **Developer Response 1:** During testing there was some missing data for 8% of exams and 90% of this was related to missing radiation dose information. **We believe that the issue of missing radiation data is for the most part entirely solvable.** The missing radiation data is not related to an entity's hardware except in very rare situations in which very old machines are used to perform the exam; rather, it is almost entirely a software and data storage issue. The radiation dose data is stored within the Radiation Dose Structured Report (RDSR), a digitized, structured summary of the total radiation output associated with the performance of the CT exam. The RDSR is produced with every CT scan and CMS incentivizes the creation of the RDSR by paying a lower reimbursement for CT scans that do not produce an RDSR. The issue that can arise is that some entities may not save and store the RDSR. There is a widespread campaign organized by the American College of Radiology to encourage entities to save and store RDSR information. Sites that do not currently save the RDSR in their radiology electronic systems will need to invest time and resources in modifying their systems to be able to do so. We calculated the amount of time this requires as part of the testing and it was quite modest, as we will describe in the Feasibility section. Although sites may require

vendor support, this work is not excessively burdensome. One of our testing sites went from saving 0% to 96% of their machines' RDSRs in a week's time with remote support from Siemens. Another site with mostly General Electric CT machines increased saving from 10% to 65% within a month, adjusting one machine at a time.

- As described in section 2b.10, **the measure steward will closely monitor missingness at the accountable entity level and report these numbers to the entities, which will be expected to fix the issue within a reasonable period of time.** If missingness doesn't resolve to near-zero by the time of NQF Maintenance, we will consider revising the measure to establish a missing data threshold beyond which exams with missing data will be treated as a failure.

Other General Comments

- **Issue 1: SOFTWARE COST.** Two reviewers inquired about the cost of the proprietary software that is required to process primary data elements from electronic systems to generate variables required by the eCQM: CT category, size-adjusted radiation dose, and global noise. The reviewers voiced concern about imposing the cost of this software package on accountable entities.
    - **Developer Response 1:** Clinicians, clinician groups, and hospitals will be able to report on the measures for free through an open web interface.