NATIONAL QUALITY FORUM
Driving measurable health improvements together

Scientific Methods Panel Discussion Guide

FALL 2022 EVALUATION CYCLE

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

https://www.qualityforum.org

Contents

Background

The Scientific Methods Panel (SMP) provides National Quality Forum (NQF) Standing Committees with evaluations of the scientific acceptability of submitted complex measures (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 2022 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 SMP, measure developers, and NQF staff. This cycle, the SMP evaluated 13 complex measures. Five are up for discussion and potential revote. One has been pulled by SMP members or NQF staff for further discussion, although they have passed NQF's Scientific Acceptability criterion. Two measures withdrew after preliminary review by the SMP. Vote totals in this discussion guide are the preliminary results and reflect votes the members were able to provide prior to the meeting. In this cycle, all measures vote totals differed between reliability, validity, and composite construction because four members were not able to submit their preliminary votes on reliability/validity/composite. The six measures that are not slated for discussion will pass with preliminary votes via consent calendar by the SMP.

After the SMP reviews measures, those that pass on scientific acceptability (either by consent calendar or by passing during the meeting) move on to their respective topic area Standing Committee for a measure evaluation of the remaining NQF standard measure evaluation criteria (i.e., Importance to Measure and Report, Feasibility, Use, Usability, and requirements for Related and Competing Measures). Measures that do not pass the SMP's review can be pulled by a Standing Committee member for further discussion and revote if it is an eligible measure. Please refer to the section titled *Scientific Methods Panel: Frequently Asked Questions* in NQF's standard measure evaluation criteria for details on this process.

Measures for Discussion (Brief)

Subgroup 1

- NQF #3725 Home Dialysis Retention (Kidney Care Quality Alliance (KCQA))
    - Reliability: H-1; M-4; L-5; I-1 Consensus Not Reached
    - Validity: H-1; M-7; L-2; I-1 Pass
- NQF #3654 Hospice Care Index (CMS/Abt Associates)
    - Reliability: H-1; M-2; L-3; I-5 No Pass
    - Validity: H-0; M-3; L-2; I-6 No Pass
    - Composite Construction: H-1; M-4; L-3; I-2 Consensus Not Reached

Subgroup 2

- NQF #3721 Patient-Reported Overall Physical Health Following Chemotherapy among Adults with Breast Cancer (Purchaser Business Group on Health)
    - Reliability: H-0; M-2; L-8; I-0 No Pass
    - Validity: H-1; M-4; L-3; I-2 Consensus Not Reached
- NQF #3720 Patient-Reported Fatigue Following Chemotherapy among Adults with Breast Cancer (Purchaser Business Group on Health)
    - Reliability: H-0; M-9; L-1; I-0 Pass
    - Validity: H-1; M-5; L-2; I-2 Consensus Not Reached
- NQF #3718 Patient-Reported Pain Interference Following Chemotherapy among Adults with Breast Cancer (Purchaser Business Group on Health)
    - Reliability: H-0; M-9; L-1; I-0 Pass
    - Validity: H-2; M-5; L-1; I-2 Pass

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

Subgroup 1

- NQF #3703 Hospitalization for Ambulatory Care Sensitive Conditions for Dual Eligible Beneficiaries enrolled in Medicare Fee-for-Service (Duals-1 FFS) or Medicare-Medicaid Plans (Duals-1 MMP) (CMS/Yale CORE)
    - Reliability: H-5; M-5; L-0; I-0 Pass
    - Validity: H-1; M-8; L-1; I-0 Pass
    - Composite Construction: H-2; M-6; L-1; I-1 Pass
- NQF #2651 CAHPS® Hospice Survey, Version 9.0 (CMS)
    - Reliability: H-6; M-3; L-2; I-0 Pass
    - Validity: H-1; M-6; L-2; I-2 Pass
- NQF #3726 Serious Illness Survey for Home-Based Programs (RAND Corporation)
    - Reliability: H-4; M-4; L-2; I-1 Pass
    - Validity: H-3; M-6; L-2; I-0 Pass
- NQF #3722 Home Dialysis Rate (KCQA)
    - Reliability: H-5; M-3; L-1; I-2 Pass
    - Validity: H-1; M-6; L-3; I-1 Pass

Subgroup 2

- NQF #2958 Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery (Massachusetts General Hospital)
    - Reliability: H-6; M-2; L-0; I-1 Pass
    - Validity: H-4; M-4; L-1; I-0 Pass
- NQF #2962 Shared Decision Making Process (Massachusetts General Hospital)
    - Reliability: H-0; M-8; L-0; I-2 Pass
    - Validity: H-3; M-4; L-1; I-2 Pass

Measures Withdrawn After SMP Review

Subgroup 1

- NQF #2881 Excess days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI) (Centers for Medicare & Medicaid Services (CMS)/Yale New Haven Health Services Corporation – Center for Outcomes Research and Evaluation (CORE))
    - Reliability: H-0; M-4; L-6; I-1 No Pass
    - Validity: H-4; M-5; L-2; I-0 Pass

Subgroup 2

- NQF #2789 Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care (Center of Excellence for Pediatric Quality Measurement)
    - Reliability: H-0; M-2; L-2; I-5 No Pass
    - Validity: H-0; M-1; L-4; I-4 No Pass

Measures for Discussion (Detailed)

Subgroup 1

NQF #3725 Home Dialysis Retention

New Measure

Brief Description of Measure: Percent of all new home dialysis patients in the measurement year for whom greater than or equal to 90 consecutive days of home dialysis was achieved.

Numerator Statement: Patients from the denominator who achieved greater than or equal to 90 consecutive days of home dialysis in the measurement year.

Denominator Statement: The total number of eligible new home dialysis patients attributed to the dialysis facility during the measurement year.

Denominator Exclusions: Denominator patients who are discharged from the facility for any of the following events occurring less than 90 days after meeting the 30-day eligibility criterion[1] are excluded:[2]

- Transplant;
- Death;
- Discontinuation of dialysis;
- Recovery of function;
- Admission to hospice; and/or
- Admission to nursing home or other LTCF.

References:

1. To account for the requisite home dialysis training period (up to 4 weeks for home hemodialysis), wherein a certain proportion of patients can be expected to drop out before completion, new home dialysis patients are not eligible for inclusion in the denominator until Day 30 following their first home dialysis treatment, at which time the consecutive time count towards the numerator criterion commences. The rationale for this "eligibility criterion" is to avoid creating a disincentive for a home dialysis trial by penalizing providers for treatment failures during this training period.
2. The exclusions are intended to avoid disincentivizing home dialysis trials by penalizing providers for unanticipated events beyond their realm of control that prevented a patient from achieving the 90 day numerator criterion.

Measure Type: Outcome: Intermediate Clinical Outcome

Data Source: Electronic Health Data, Electronic Health Records

Level of Analysis: Facility

Risk-Adjusted: Stratification by five risk factor groups

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: Consensus was not reached by the SMP on Reliability with a score of: H-1; M-4; L-5; I-1

Specifications:

- This measure was previously submitted to the SMP under NQF #3697 as a clinical intermediate outcome measure. The developer has resubmitted it as a clinical intermediate outcome measure under NQF #3725.
- To account for previous feedback from the SMP, the developer kept the level of analysis at the facility but provided an explanation as to why HRR-level analysis is not required (see below under "Reliability Testing").
- Measure specifications are clear and precise.

Reliability Testing:

- Reliability testing was conducted at the accountable-entity level:
    - Reliability testing was conducted at the facility level using signal-to-noise analysis: the beta-binomial model. The developer states that HRR-level aggregation is not necessary for this measure because it only includes incident patients and does not need to account for facilities that do not offer home dialysis.
    - Mean reliability at the facility level (N=2,812) using one year of data was 0.604 (median= 0.547). The median facility had seven patients.
    - The developer noted that while the reliability statistics using one year of data meet NQF's criteria, they also calculated reliability by duplicating their data and treating it as a two-year rolling measure, given the small numbers of new home dialysis patients. The mean reliability increased to 0.846 (a median of 0.905) with the second year of data.
    - The developers noted that in order to confirm that the double use of the 2021 data provides a valid analysis, they performed an additional analysis by randomly generating new yearly data for each facility and combined that with the 2021 data, resulting in a similar increase in reliability (0.871 with a median of 0.931). The developers argue that this additional analysis helps to alleviate concerns of auto-correlation.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-1; M-7; L-2; I-1

Validity Testing

- Validity testing was conducted at the accountable-entity level:
    - Validity testing was conducted using face validity with a panel of nine members (five healthcare providers, two dialysis facilities, and three manufacturer groups).
    - Seven of the nine members agreed that the measure score likely or highly likely provides an accurate reflection of quality and that the measure would effectively distinguish real differences in performance between providers.
    - Eight of the nine members agreed that the measure scores for the paired set (NQF #3722 and NQF #3725) will provide an accurate reflection of quality and that the paired set will effectively distinguish real differences in performance between providers.
    - A dissenting member noted concerns about the minimal patient exclusion criteria and that this would make the measure more of a reflection of the provider's patient population and not their performance.

Exclusions

- The following exclusions are applied to the denominator: patient months with hospice (0.0 percent); patient months in a nursing home or other LTCF (1.0 percent); and patient discharge secondary to transplant (0.4 percent), death (1.8 percent), discontinuation of dialysis (0.5 percent), and/or recovery of renal function in the month (0.1 percent). After accounting for overlap in exclusions, a total of 3.1 percent unique patient months were excluded from the denominator. The mean facility-level performance before exclusions was 72.4 percent, and with them applied, it was 74.7 percent. The developer notes that these exclusions are clinically warranted to avoid creating a disincentive for home dialysis trials by penalizing providers for unanticipated events beyond their control that prevent a patient from achieving the 90-day numerator criterion.

Risk Adjustment

- The developer risk-stratified the measure by age, gender, race/ethnicity, and dual-eligible status. They also explored markers of functional risk and clinical variables for stratification, but they were not included due to data availability.
- Stratified analyses of performance demonstrate that a clear trend by age (with patients under the age of 18 achieving 90 or more days of home dialysis more consistently than older age groups), differences by race (with higher performance in "Other" races than in Black or White patients) and ethnicity (with Hispanics performing more than 7 percent higher than non-Hispanics), and by insurance status (with dual-eligible patients performing slightly better than non–dual-eligible patients).

Meaningful Differences

- The mean performance was 74.7 percent and the 25th percentile. The median and 75th percentile performance scores are 69 percent, 83.3 percent, and 100 percent.
- To demonstrate the statistical significance of the spread, the developer used the 2021 data and the randomly generated data and analyzed 1,699 facilities with a non-zero performance score. The overall weighted mean performance score was 80.4 percent with the facility size as the weight. The developer noted this as the national norm. Sixty percent of facilities with a score between 6.25 percent and 52.87 percent (below the 10th decile) had 95 percent CIs below the norm. Facilities with a score greater than 92.86–98.53 percent (90th decile and above) all had 95 percent CIs above the norm. The developers noted that measure performance scores can identify facilities with good performance, but the identification of facilities with poor performance was more variable likely due to the small facility size.

Missing Data

- The developer notes that while they believe their observed percent of patient-months excluded secondary to hospice enrollment is not accurate, they believe those same patients are captured in other exclusions. The developer also believes their observed percent of patient-months excluded due to nursing/LTCF residence is an underestimate. However, they note that if they were to use the highest exclusion rate reported, there is only a difference of 0.4 percent in the overall facility-level score.
- When patient months were excluded from the denominator due to missing values in the stratification variables (e.g., age, sex, race, ethnicity, and dual-eligibility status), the mean facility-level performance was 74.7 percent before exclusions and 74.8 percent after excluding missing values.

Comparability

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

SMP Concerns

- Two SMP members noted that the Standing Committee should discuss what happens if a dialysis patient enters the denominator after October 1 and cannot meet the 90 day threshold and whether the choice of only including patients already retained for 30 days is best.
- The measure is specified for one year, but the measure developer advised that if the measure is implemented, reliability would be improved with a two-year construct. One SMP member noted that, given the developers response to NQF clarification on the two-year rolling requirement, the measure should not be implemented as a one-year measure. Several other members noted that low volume units do not have adequate reliability for the one-year measure as well. Lastly, one SMP member noted that the calculations used for reliability may be overestimating the true reliability due to the small facility-specific denominators and a lack of precision in the denominator estimates.

Items to be Discussed

- Discuss and revote on reliability as it received a consensus not reached rating. Votes of low and insufficient were due to the low volume units not obtaining adequate reliability using one-year of data, as the measure is specified, as well as concerns surrounding the calculation used for the reliability score.

NQF #3654 Hospice Care Index

New Measure

Brief Description of Measure: The Hospice Care Index (HCI) monitors a broad set of leading, claims-based indicators of hospice care processes. It reflects care throughout the hospice stay and by the care team within the domains of higher levels of care, visits by nursing staff, patterns of live discharge, and per-beneficiary spending. The index monitors ten indicators simultaneously, and compares individual provider scores to the thresholds which are set as benchmarks against the national distribution of performance scores. Hospices which are not outliers (comprising the vast majority of the range of scores) are awarded a point for that indicator.

The index is calculated as follows: across the ten indicators, the measure flags hospices with the most extreme scores defined as surpassing a particular threshold; e.g., "the bottom 10% of hospices by nursing minutes per day". All hospices that do not exceed a threshold are assigned a point for that indicator. The measure then assigns the hospices an overall index score calculated as the total of all assigned points for the indicators; i.e., if a hospice never crossed a threshold, it would receive a point for all ten indicators, and its score would be 10. If it crossed one threshold, it would only be given a point for nine of the ten indicators, and its score would be a 9. The HCI's total score ranges from a 0-10, where a perfect "10" indicators the hospice was not an outlier for any indicator and has performance scores commensurate with the vast majority of nationwide providers. Because its indicators include a multitude of topics, and hospices receive lower scores if they display as outliers across multiple indicators, the HCI seeks to identify hospices which are outliers across an array of different areas of hospice care, simultaneously.

The index feature of this measure – covering multiple topics at once – was intentional to respond to comments received during Federal Fiscal Year 2020 rulemaking (84 FR 38484; https://www.govinfo.gov/content/pkg/FR-2019-08-06/pdf/2019-16583.pdf) that expressed concern about a previously developed live discharge/transitions from hospice measure. Commenters expressed

that there were limits to what a single claims-based measure (of hospice transitions) could convey; i.e., that there could be other explanations for a hospice's poor performance than the claims information convey. Again, by identifying hospices which are outliers in multiple areas simultaneously, this index assigns hospices as outliers with more reliability and internal validity than a single-outcome claims measures and thereby overcomes its limitations. The index's focus on outliers acknowledges that some prevalence of the indicators is normal and is expected.

More broadly, the Hospice Care Index monitors the performance for a broad and holistic set of indicators for hospice care processes not otherwise addressed within the current quality measures of CMS's Quality Reporting Program.

The topics which the indicators capture were taken from a review of recommendations and reports by the Office of Inspector General (OIG, reports by MedPAC, or in the academic literature. These domains include the provision of hospice services, live discharges, and levels of care. The HCI will add value to the Hospice Quality Reporting Program (HQRP) by addressing topics beyond the current measure set. Because, this measure is calculated using administrative records only, it will provide information to the public with no additional burden to patients, their caregivers, or hospices.

Each indicator is a key component of the HCI measure, and the ten indicators assessed in the HCI are described in the table below followed by indicator specifications.

Numerator Statement: This index numerator is based on an approved NQF approach and does not have a traditional numerator. The index score is calculated as the total number of points each across the ten hospice-level indicators. Hospices earn a point on each indicator if their indicator scores do not cross an assigned threshold for that indicator. Then, the overall index score is calculated as the total sum of points across the ten indicators. Therefore, the potential range of scores is from 0 (earning no points) to 10 (a perfect score, where a point is earned for each indicator).

Denominator Statement: The Hospice Care Index does not have a traditional numerator; the index is scored as the number of ten indicators by which the hospice earns a point, based on their performance in each indicator; i.e., the "numerator" is a score of 0 to 10.

The ten indicators that comprise the composite do have their own numerator statements; technical specifications for each are detailed in sp.22).

Denominator Exclusions: There are no exclusions based on types of hospices. However, the measure steward (CMS) maintains a minimum public reporting threshold of at least 20 quality episodes to ensure reliable provider-level results, and measure testing was only conducted on hospices meeting this threshold so that testing results would be aligned with what is actually publicly reported (in the case of the HCI, hospices needed to have at least 20 claims to be included in testing). However, we consider this more a note of the composition of the analytic file than a specification exclusion, per se. Note also that the index was developed using claims from Federal Fiscal Years 2018 through 2019.

Measure Type: Composite

Data Source: Claims

Level of Analysis: Facility

Not Risk-Adjusted

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Did Not Pass on Reliability with a score of: H-1; M-2; L-3; I-5

Specifications:

- Measure specifications are clear and precise.
- Measure specifications for the composite performance measure also include component measure specifications, aggregation and weighting rules, and required sample sizes.

Reliability Testing:

- Reliability testing was conducted at the patient/encounter level:
  - No patient/encounter level testing was provided.
- Reliability testing was conducted at the accountable-entity level:
  - The developer indicates that the traditional approach of signal-to-noise ratio (SNR) testing as outlined in the Adams (2009) tutorial is not applicable to this measure.
  - The developer instead conducts a stability analysis, comparing index scores calculated for the same hospice (n=3,576) using FFY 2017 and FFY 2019. No statistical tests were conducted.
    - Forty-six percent had the same score in 2017 and 2019.
    - Fifteen percent had scores that differed by two points or more.
  - The developer states that the design of the index, with its focus on identifying hospices that are outliers in several areas, ensures its reliability.

Validity

Preliminary ratings for validity: The SMP Did Not Pass on Validity with a score of: H-0; M-3; L-2; I-6

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - No patient/encounter level testing was provided.
- Validity testing was conducted at the accountable-entity level:
  - The developer compared the HCI score to the following: (1) the percent of caregivers rating the hospice nice or 10 (out of 10) and (2) the percent of caregivers that would "definitely" recommend the hospice. Pearson's correlation coefficient was calculated (0.0916 and 0.1155, respectively, $p<0.001$).
  - The developer estimated a simple logistic regression estimated with an HCI score as an explanatory variable and CAHPS' Star Rating (summarizing hospice ratings across eight CAHPS hospice outcomes in a single score). The odds ratio was 2.02 (95% Confidence Interval [CI] of 1.08–3.79). A hospice with a HCI score of seven or below is twice as likely to receive the lowest Star Ratings compared with a hospice that has an index score of 10.

Exclusions

- The measure does not have any exclusions.

Risk Adjustment

- The measure is not risk-adjusted or stratified.
- The developer notes that the HCI is a composite measure of hospice processes; therefore, no risk adjustment or risk stratification is used.

Meaningful Differences

- The developer calculated the percentage of hospices achieving each of the 11 possible scores (0 through 10). Scores ranged from 2–10.
- Seventy-one percent of hospices had HCI scores of 9 or 10.
- A total of 12.6 percent of hospices had HCI scores below 8.

Missing Data

- The developer notes that the measure is based on Medicare claims, "which are considered a complete data source."

Comparability

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

Composite

Preliminary ratings for composite construct: Consensus was Not Reached by the SMP on composite with a score of: H-1; M-4; L-3; I-2

Empirical analysis to support composite construction

- Face validity assessments were used to develop the composite.
- Components of the composite were identified in the public recommendations to CMS from other federal agencies.
- An environmental scan was conducted to identify component domains and indicators, and a TEP was also engaged. The developer provided a link to the TEP composition and a report describing the TEP discussions.
- The developer conducted an iterative simulation of removing each index indicator in turn and recalculating the performance scores and standard deviations of the composite. The standard deviation with all indicators included was 1.200. The standard deviations with each indicator in turn removed ranged from 1.073–1.1172.

SMP Concerns

- Some SMP members expressed concerns with the clarity of the specifications noting that the submission says that there are no exclusions by facility type but subsequently says facilities with fewer than 20 episodes are excluded; it is unclear if sampling was done, if the term outlier refers to both high and low outliers or just high. Generally, reviewers found the specifications complex and were concerned that they may not be reproducible upon implementation. Finally, although the data elements are based on a location in a distribution, information on the actual distributions are not provided. Small differences between low and high values suggest that some scoring is arbitrary or random, but this cannot be assessed from the information provided.
- The SMP noted concerns regarding the reliability testing. A number of members sated that the stability analysis presented may not have been sufficient. Several SMP members noted that while the developer's assertion is correct that a signal-to-noise test was not possible, they could have performed a test-retest analysis or data element testing and therefore the testing provided was insufficient.
- Most SMP members were concerned that the Pearson correlations in the validity testing results were too low to show validity. Members noted that the face validity testing results would best be assessed by the Standing Committee.

- Several SMP members noted concerns with the analysis done for meaningful differences noting that because the developer does not provide empirical evidence on how point differences should be interpreted, it is not clear that there are meaningful differences.
- While the measure is not risk adjusted, some SMP members noted concerns that the measure may require adjustment and this should be discussed by the Standing Committee. These reviewers questioned that, given the assumption that if some of the items were standalone measures they would be considered outcomes and thus would be risk adjusted. They concluded that the rationale for not adjusting the larger composite was insufficient.
- Some SMP members noted concern with the composite construction noting that information was insufficient to assess the construction and voiced concerns with the reliance on a TEP to assess composite construction and the limited empirical analysis.

Items to be Discussed

- Discuss the developer's response to the SMP's concerns with the methodologies and results for the reliability and validity testing.
- Discuss and revote on the measure's composite construction as it received a consensus not reached rating due to the lack of information provided to assess the construction.

Subgroup 2

NQF #3721 Patient-Reported Overall Physical Health Following Chemotherapy Among Adults With Breast Cancer

New Measure

Brief Description of Measure: The PRO-PM assesses overall physical health among adult women with breast cancer entering survivorship after completion of chemotherapy administered with curative intent. Overall physical health is assessed using the PROMIS Global Health v1.2 scale administered at baseline (prior to chemotherapy) and at follow-up (about three months following completion of chemotherapy). The measure is risk-adjusted.

Numerator Statement: The PRO-PM numerator is the group-level PROMIS Overall Physical Health score at the follow-up survey.

Denominator Statement: Adult patients with stages I-III female breast cancer receiving an initial chemotherapy regimen within the measurement window.

Denominator Exclusions:

- Patients on a therapeutic clinical trial
- Patients with recurrence/disease progression
- Patients who leave the practice
- Patients who die

Measure Type: Outcome: PRO-PM

Data Source: Electronic Health Records, Instrument-Based Data, Paper Medical Records

Level of Analysis: Clinician: Group/Practice

Risk-Adjusted: Statistical risk model with 13 factors

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Did Not Pass on Reliability with a score of: H-0; M-2; L-8; I-0

Specifications:

- The PRO-PM is the risk-adjusted group-level mean of Patient-Reported Outcomes Measurement Information System (PROMIS) overall physical health scores among adult women with breast cancer entering survivorship after the completion of chemotherapy administered with curative intent.
- Measure specifications are clear and precise.
- Measure specifications for this instrument-based measure also include the specific instrument (e.g., PROM[s]) and standard methods, modes, and languages of administration.

Reliability Testing:

- Data were used from 7/1/19 to 4/1/22 at 10 group practices.
- Reliability testing was conducted at the encounter level and accountable-entity level.

- The developer notes that PROMIS measures, including the overall physical health scale, have undergone rigorous development and validation. Several references are provided in the submission.
    - Reliability testing from the literature demonstrates that for the PROMIS Global Health, the Cronbach's alphas are 0.92 (overall), 0.81 (physical health) and 0.86 (mental health).
- To test the reliability of the measure score, a signal-to-noise analysis was performed. To evaluate measure reliability for group-level reporting, hierarchical linear regression models were used to relate the outcome to providers and covariates. The hierarchy was patients' observations within groups.
    - The estimate of the adjusted ICC was 0.034. The estimate of the reliability at the average sample size for a group (32 patients per group) was 0.534.
    - Using the Spearman-Brown prophecy formula, the developer estimates that in order to obtain a nominal reliability of 0.7, a minimum sample size of 66 patient respondents would be required. Group specific reliability ranged from 0.18–0.70, with a mean of 0.45 (SD=0.23) and a median reliability of 0.44.
    - The proportion of groups in the sample that had sufficient reliability using a reliability threshold of 0.70 was 10 percent.

Validity

Preliminary ratings for validity: The SMP Did Not Reach Consensus on Validity with a score of: H-1; M-4; L-3; I-2

Validity Testing

- Validity testing was conducted at the encounter level for critical data elements using the PROMOnc data registry. The developer stated that the majority of clinical and demographic variables could be validated, but several variables were excluded from testing because they were not in the registry used for the validity testing. Details on variables excluded are on page 24 of the submission.
    - Five hundred seventy patients were included in this analysis.
    - The percentage agreement by data element ranged from 71.63–100 percent. Reported kappas ranged from 0.64–0.67. The reported sensitivity ranged from 33.33–89.52 percent. The specificity ranged from 60–99.45 percent. The data can be found in Table 2b.1. Several cells in this table were intentionally left blank.
- Validity testing of the measure score was conducted through a systematic assessment of face validity using a panel of 12 oncologist advisors. The following survey question was asked: "Rate your agreement with the following statement: The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."
    - Eight of the 12 advisors participated in the survey.
    - All eight indicated "moderate agreement," "agreement," or "strong agreement" to the above question (3, 4, or 5 out of 5).
    - For physical health, all eight agreed or strongly agreed (4 or 5 out of 5) that the measure could differentiate good versus poor quality.

- The four oncologists who declined to participate in the face validity voting expressed concerns regarding the impact of COVID-19 on sample size, and thus, performance scores. They requested additional data prior to voting.

Exclusions

- There were no concerns about the exclusion analysis.
- There are four exclusions (n=frequency of those exclusions from the measure denominator):
  - Patients on an interventional or therapeutic clinical trial (n=18)
  - Patients who experience relapse or disease progression (n=0)
  - Patients who leave the practice (n=0)
  - Patients who die (n=1)

Risk Adjustment

- A statistical model is used to risk-adjust this measure using 13 variables.
- To estimate risk-adjusted quality measure scores, hierarchical linear models that relate the patient-measure score to group scores conditioned on risk adjustment covariates were used.
- Model discrimination was tested during the Kendall tau. Comparing scores between null and the multivariate model adjustments for pain interference resulted in a value of 0.56. The Pearson correlation coefficient between the observed and predicted responses was 0.50.

Meaningful Differences

- There are some concerns about meaningful differences, as only one in 10 of the sites were different from the mean, compared to two out of 10 for the pain interference and fatigue scales.
- To examine the ability of the measure to identify high- or low-performing groups, the developer calculated the number and percentage of groups that were significantly above or below the average score using risk adjustment.
- The mean group performance score was 44.76 and the standard deviation was 2.63, with a median score of 44.36 and a range of 40.34 to 49.88. One of 10 groups had significantly different scores than the overall average, being less favorable. The mean absolute difference between the group's scores and the overall average was 5.19 points on a T-score scale (SD=10).

Missing Data

- There are no large concerns about missing data.
- Both survey nonresponse and missing data were assessed.
- Across the 10 sites, 896 patients were eligible for follow-up and 19 met the exclusion criteria. The total number of follow-up surveys was 744, making up a survey administration rate of 85 percent. Among those surveys, 323 were completed and nine were ineligible. No statistical significance was identified, except that the respondents and nonrespondents differed on marital status and insurance.
- Missingness ranged from 0.31–0.93 percent for PROMIS item scales.

Comparability

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

SMP Concerns

- As with #3718 and 3720, several SMP members had concerns with the specifications around collecting both baseline and follow-up surveys. It was unclear to members why collection of both was necessary and which were used in the calculation of the measure. One SMP member was concerned that the calculation of the adjusted scores was not explicit enough. Specifically, how the group practice effect was accounted for in the calculation. There were concerns that oral chemotherapy, because the allowable response window includes 7 days after oral chemotherapy start date, could be problematic as some patients may already have started to experience the effects of chemotherapy. Therefore, baseline survey scores might be affected. Because baseline scores were used for risk adjustment, this could lead to biased results.
- Unlike the other measures in this group, there were significant concerns with the accountable entity level reliability testing results as only 1 of the 10 groups involved in testing had enough patients to reach a reliability score of 0.7. It was noted that the reliability is 0.45 for physical health and 0.44 for mental health. Ultimately, SMP did not pass the measure on reliability based on these concerns.
- In the data element validity testing, it was noted by some SMP members that some results, particularly the sensitivity, were low. Agreement ranged 71.63 - 100 percent, sensitivity ranged from 33.33 - 89.52 percent, and specificity ranged from 60 - 99.45 percent.
- As with #3718 and 3720, SMP members had mixed reviews on the accountable entity level validity testing, specifically, the face validity testing. Some members viewed the results as adequate as 8/12 experts agreed or strongly agreed that this measure differentiated quality. However, others were concerned with the four non-respondents and the lack of detail regarding the selection process. One member voiced a specific concern that there were no patients present on the TEP that evaluated face validity, but recognized that several TEP members did not respond or vote. Another member noted that the theory of quality for the measure is unclear and this further weakens the face validity results.
- As with #3718 and 3720, several reviewers did not find the testing to demonstrate meaningful differences adequate. One member noted that only 10 sites were in the testing sample, while others noted that in that small sample, only one was found to have significantly different scores from the overall average.
- As with #3718 and 3720, one member requested additional detail regarding missing response rates by site, not just an overall level of missingness. This reviewer also noted that some calculations overestimated the missingness.
- Unlike the other measures in this group, SMP members also voiced concerns with the testing results of the risk adjustment model. Specifically, the following results: Kendall's Tau = 0.56 and Pearson correlation = 0.50.

Items to be Discussed

- Discuss the developer's response to the SMP's concerns with the reliability and validity testing.
- Discuss and revote on validity as it received a consensus not reached rating. Votes of low and insufficient were due to concerns with the face validity testing, the lack of demonstration of meaningful differences, and missing response rates.
- This measure is grouped with #3720 and #3721. The SMP should discuss whether all testing results are different enough to warrant different votes on the three measures.

NQF #3720 Patient-Reported Fatigue Following Chemotherapy Among Adults With Breast Cancer

New Measure

Brief Description of Measure: The PRO-PM assesses fatigue among adult women with breast cancer entering survivorship after completion of chemotherapy administered with curative intent. Fatigue is assessed using the PROMIS Fatigue 4a scale administered at baseline (prior to chemotherapy) and at follow-up (about three months following completion of chemotherapy). The measure is risk-adjusted.

Numerator Statement: The PRO-PM numerator is the group-level PROMIS Fatigue score at the follow-up survey.

Denominator Statement: Adult patients with stages I-III female breast cancer receiving an initial chemotherapy regimen within the measurement window.

Denominator Exclusions:

- Patients on a therapeutic clinical trial
- Patients with recurrence/disease progression
- Patients who leave the practice
- Patients who die

Measure Type: Outcome: PRO-PM

Data Source: Electronic Health Records**,** Instrument-Based Data**,** Paper Medical Records

Level of Analysis: Group/Practice

Risk-Adjusted: Statistical risk model with 13 factors

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-0; M-9; L-1; I-0

Specifications:

- The PRO-PM is the risk-adjusted, group-level mean of PROMIS Fatigue scores among adult women with breast cancer entering survivorship after the completion of chemotherapy administered with curative intent.
- Measure specifications are clear and precise.
- Measure specifications for this instrument-based measure also include the specific instrument (e.g., PROM[s]) and standard methods, modes, and languages of administration.

Reliability Testing:

- Data were used from 7/1/19 to 4/1/22 at 10 group practices.
- Reliability testing was conducted at the encounter level and accountable-entity level.
- The developer notes that PROMIS measures, including the Fatigue scale, have undergone rigorous development and validation. Several references are provided in the submission.
    - Reliability testing from the literature demonstrates that for the PROMIS Fatigue, the Cronbach's alpha is 0.86.

- To test the reliability of the measure score, a signal-to-noise analysis was performed. To evaluate the measure's reliability for group-level reporting, hierarchical linear regression models were used to relate the outcome to providers and covariates. The hierarchy was patients' observations within groups.
    - The estimate of the adjusted ICC was 0.094. The estimate of the reliability at the average sample size for a group (32 patients per group) was 0.77.
    - Using the Spearman-Brown prophecy formula, the developer estimates that in order to obtain a nominal reliability of 0.7, a minimum sample size of 23 patient respondents would be required. Group specific reliability ranged from 0.38 to 0.88, with a mean of 0.66 (SD=0.21) and a median reliability of 0.68.
    - The proportion of groups in the sample that had sufficient reliability using a reliability threshold of 0.70 was 50 percent.

Validity

Preliminary ratings for validity: The SMP Did Not Reach Consensus on Validity with a score of: H-1; M-5; L-2; I-2

Validity Testing

- Validity testing was conducted at the encounter level for critical data elements using the PROMOnc data registry. The developer stated that the majority of the clinical and demographic variables could be validated, but several variables were excluded from testing because they were not in the registry used for the validity testing. Details on excluded variables are on page 24 of the submission.
    - Five hundred seventy patients were included in this analysis.
    - The percentage agreement by data element ranged from 71.63–100 percent. Reported kappas ranged from 0.64–0.67. The reported sensitivity ranged from 33.33–89.52 percent. The specificity ranged from 60–99.45 percent.
- Validity testing of the measure score was conducted through a systematic assessment of face validity using a panel of 12 oncologist advisors. The following survey question was asked: "Rate your agreement with the following statement: The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."
    - Eight of the 12 advisors participated in the survey.
    - All eight indicated "moderate agreement," "agreement," or "strong agreement" to the above question.
    - Three agreed or strongly agreed that the fatigue measure could differentiate good versus poor quality. Participants who did not rate the measure as 4 or 5 (i.e., agree or strongly agree) felt that fatigue was more susceptible to pandemic-related issues.
    - The four oncologists who declined to participate in the face validity voting expressed concerns regarding the impact of COVID-19 on sample size, and thus, performance scores. They requested additional data prior to voting.

Exclusions

- There are no concerns about the exclusion analysis.
- There are four exclusions (n=frequency of those exclusions from the measure denominator):
    - Patients on an interventional or therapeutic clinical trial (n=18)

- Patients who experience relapse or disease progression (n=0)
- Patients who leave the practice (n=0)
- Patients who die (n=1)

Risk Adjustment

- A statistical model is used to risk-adjust this measure using 13 variables.
- To estimate risk-adjusted quality measure scores, hierarchical linear models that relate the patient-measure score to group scores conditioned on risk adjustment covariates were used.
- Model discrimination was tested during the Kendall tau. Comparing scores between null and the multivariate model adjustments for pain interference resulted in a value of 0.87. The Pearson correlation coefficient between the observed and predicted responses was 0.55.

Meaningful Differences

- There were no concerns about meaningful differences, considering below 20 percent of the sites were different from the mean.
- To examine the ability of the measure to identify high- or low-performing groups, the developer calculated the number and percentage of groups that were significantly above or below the average score using risk adjustment.
- The mean group performance score was 48.51, and the standard deviation was 3.13, with a median score of 48.67 and a range of 42.13–53.07. Two of 10 groups had significantly different scores than the overall average, one more favorable and the other less favorable. Among those two groups, the mean absolute difference between the group's scores and the overall average was 4.9 points on a T-score scale (SD=10).

Missing Data

- There are no large concerns about missing data.
- Both survey nonresponse and missing data were assessed.
- Across the 10 sites, 896 patients were eligible for follow-up and 19 met the exclusion criteria. The total number of follow-up surveys was 744, making up a survey administration rate of 85 percent. Among those surveys, 323 were completed and nine were ineligible. No statistical significance was identified, except that the respondents and nonrespondents differed on marital status and insurance.
- Missingness ranged from 0.00–0.93 percent for PROMIS item scales.

Comparability

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

SMP Concerns

- As with #3718 and #3721, several SMP members had concerns with the specifications around collecting both baseline and follow-up surveys. It was unclear to members why collection of both was necessary and which were used in the calculation of the measure. One SMP member was concerned that the calculation of the adjusted scores was not explicit enough. Specifically, how the group practice effect was accounted for in the calculation. There were concerns that oral chemotherapy, because the allowable response window includes 7 days after oral chemotherapy start date, could be problematic as some patients may already have started to experience the effects of chemotherapy. Therefore, baseline survey scores might be affected. Because baseline scores were used for risk adjustment, this could lead to biased results.

- Some members viewed the accountable entity level face validity testing results as adequate, but unlike #3718 and #3721, a greater number of SMP members were concerned with the low results compared to the other measures in this set. This was often noted in the question regarding whether the measure can differentiate good vs poor quality (only 3 agreed or strongly agreed). Again, the four non-respondents and the lack of detail regarding the selection process as well as the lack of patients present on the TEP were at issue. Again, one member noted that the theory of quality for the measure is unclear and this further weakens the face validity results.
- As with #3718, several reviewers did not find the testing to demonstrate meaningful differences adequate. One member noted that only 10 sites were in the testing sample, while others noted that in that small sample, only two were found to have significantly different scores from the overall average.
- As with #3718, one member requested additional detail regarding missing response rates by site, not just an overall level of missingness. This reviewer also noted that some calculations overestimated the missingness.

Items to be Discussed

- The SMP should discuss and revote on the reliability criterion.
- Discuss and revote on validity as it received a consensus not reached rating. Votes of low and insufficient were due to concerns with the face validity testing, the lack of demonstration of meaningful differences, and missing response rates.
- This measure is grouped with #3720 and #3721. The SMP should discuss whether testing results (especially reliability testing results) are different enough to warrant different votes on the three measures.

NQF #3718 Patient-Reported Pain Interference Following Chemotherapy Among Adults With Breast Cancer

New Measure

Brief Description of Measure: The PRO-PM assesses pain interference among adult women with breast cancer entering survivorship after completion of chemotherapy administered with curative intent. Pain interference is assessed using the PROMIS Pain Interference 4a scale administered at baseline (prior to chemotherapy) and at follow-up (about three months following completion of chemotherapy). The measure is risk-adjusted.

Numerator Statement: The PRO-PM numerator is the group-level PROMIS Pain Interference score at the follow-up survey.

Denominator Statement: Adult patients with stages I-III female breast cancer receiving an initial chemotherapy regimen within the measurement window.

Denominator Exclusions:

- Patients on a therapeutic clinical trial
- Patients with recurrence/disease progression
- Patients who leave the practice
- Patients who die

Measure Type: Outcome: PRO-PM

Data Source: Electronic Health Records, Paper Medical Records, Instrument-Based Data

Level of Analysis: Clinician: Group/Practice

Risk-Adjusted: Statistical risk model with 13 factors

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-0; M-9; L-1; I-0

Specifications:

- The PRO-PM is the risk-adjusted, group-level mean of PROMIS Pain Interference scores among adult women with breast cancer entering survivorship after the completion of chemotherapy administered with curative intent.
- Measure specifications are clear and precise.
- Measure specifications for this instrument-based measure also include the specific instrument (e.g., PROM[s]) and standard methods, modes, and languages of administration.

Reliability Testing:

- Data were used from 7/1/19 to 4/1/22 at 10 group practices.
- Reliability testing was conducted at the encounter level and accountable-entity level.
- The developer notes that PROMIS measures, including the pain interference scale, have undergone rigorous development and validation. Several references are provided in the submission.
    - Reliability testing from the literature demonstrates that for the PROMIS Pain interference, the Cronbach's alpha is 0.99.
- To test the reliability of the measure score, a signal-to-noise analysis was performed. To evaluate measure reliability for group-level reporting, hierarchical linear regression models were used to relate the outcome to providers and covariates. The hierarchy was patients observations' within groups.
    - The estimate of the adjusted ICC was 0.097. The estimate of the reliability at the average sample size for a group (32 patients per group) was 0.77.
    - Using the Spearman-Brown prophecy formula, the developer estimates that in order to obtain a nominal reliability of 0.7, a minimum sample size of 22 patient respondents would be required. Group specific reliability ranged from 0.39–0.88, with a mean of 0.66 (SD=0.20) and a median reliability of 0.68.
    - The proportion of groups in the sample that had sufficient reliability using a reliability threshold of 0.70 was 50 percent.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-2; M-5; L-1; I-2

Validity Testing

- Validity testing was conducted at the encounter level for critical data elements using the Patient-Reported Outcomes in Oncology (PROMOnc) data registry. The developer stated that the majority of clinical and demographic variables could be validated, but several variables were

excluded from testing because they were not in the registry used for the validity testing. Details on variables excluded are on Page 24 of the submission.
    - Five hundred seventy patients were included in this analysis.
    - The percentage agreement by data element ranged from 71.63–100 percent. Reported kappas ranged from 0.64–0.67. Reported sensitivity ranged from 33.33–89.52 percent. Specificity ranged from 60–99.45 percent.
- Validity testing was conducted at the accountable-entity level through a systematic assessment of face validity using a panel of 12 oncologist advisors. The following survey question was asked: "Rate your agreement with the following statement: The scores obtained from the measure as specified will provide an accurate reflection of quality and can be used to distinguish good and poor quality."
    - Eight of the 12 advisors participated in the survey.
    - All eight indicated "moderate agreement," "agreement," or "strong agreement" to the above survey question (i.e., 3, 4, or 5 out of 5).
    - Seven agreed or strongly agreed (i.e., 4 or 5 out of 5) that the pain interference measure could differentiate good versus poor quality.
    - The four oncologists who declined to participate in the face validity voting expressed concerns regarding the impact of coronavirus disease 2019 (COVID-19) on sample size, and thus, performance scores. They requested additional data prior to voting.

Exclusions

- There are no concerns about the exclusion analysis.
- There are four exclusions (n=frequency of those exclusions from the measure denominator):
    - Patients on an interventional or therapeutic clinical trial (n=18)
    - Patients who experience relapse or disease progression (n=0)
    - Patients who leave the practice (n=0)
    - Patients who die (n=1)

Risk Adjustment

- A statistical model is used to risk-adjust this measure using 13 variables.
- To estimate risk-adjusted quality measure scores, hierarchical linear models that relate the patient-measure score to group scores conditioned on risk adjustment covariates were used.
- The regression coefficients are described in Table 2b.3 of the measure submission form.
- Model discrimination was tested during the Kendall tau. Comparing scores between null and the multivariate model adjustments for pain interference resulted in a value of 0.64. The Pearson correlation coefficient between the observed and predicted responses was 0.53.

Meaningful Differences

- There are no concerns about meaningful differences. Twenty percent of sites were different from the mean.
- To examine the ability of the measure to identify high- or low-performing groups, the developer calculated the number and percentage of groups that were significantly above or below the average score using risk adjustment.
- The mean group performance score was 50.51, and the standard deviation was 2.83, with a median score of 50.75 and a range of 43.92–54.11. Two of 10 groups had significantly different

scores than the overall average, one more favorable and the other less favorable. Among those two groups, the mean absolute difference between the group's scores and the overall average was 4.26 points on a T-score scale (SD=10).

Missing Data

- There are no large concerns about missing data.
- Both survey nonresponse and missing data were assessed.
- Across the 10 sites, 896 patients were eligible for follow-up and 19 met the exclusion criteria. The total number of follow-up surveys was 744, making up a survey administration rate of 85 percent. Among those surveys, 323 were completed and nine were ineligible. No statistical significance was identified, except that the respondents and nonrespondents differed on marital status and insurance.
- The missingness ranged from 0.93–3.10 percent for PROMIS item scales.

Comparability

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

SMP Concerns

- As with #3720 and #3721, several SMP members had concerns with the specifications around collecting both baseline and follow-up surveys. It was unclear to members why collection of both was necessary and which were used in the calculation of the measure. One SMP member was concerned that the calculation of the adjusted scores was not explicit enough. Specifically, how the group practice effect was accounted for in the calculation. There were concerns that oral chemotherapy, because the allowable response window includes 7 days after oral chemotherapy start date, could be problematic as some patients may already have started to experience the effects of chemotherapy. Therefore, baseline survey scores might be affected. Because baseline scores were used for risk adjustment, this could lead to biased results.
- There were mixed reviews of whether the reliability testing showed completely sufficient levels of reliability. Some reviewers noted that the results were high to moderate, while others noted that the results were moderate to low.
- As with #3721, SMP members also had mixed reviews on the accountable entity level validity testing, specifically, the face validity testing. Some members viewed the results as adequate but others were concerned with the non-respondents and the lack of detail regarding the selection process. One member voiced a specific concern that there were no patients present on the TEP that evaluated face validity, but recognized that several TEP members did not respond or vote. Another member noted that the theory of quality for the measure is unclear and this further weakens the face validity results.
- As with #3720 and #3721, several reviewers did not find the testing to demonstrate meaningful differences adequate. One member noted that only 10 sites were in the testing sample, while others noted that in that small sample, only two were found to have significantly different scores from the overall average.
- As with #3720 and 3721, one member requested additional detail regarding missing response rates by site, not just an overall level of missingness. This reviewer also noted that some calculations overestimated the missingness.

Items to be Discussed

- This measure is grouped with #3720 and #3721. The SMP should discuss whether all testing results are different enough to warrant different votes on the three measures.

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

Subgroup 1

NQF #3703 Hospitalization for Ambulatory Care Sensitive Conditions for Dual-Eligible Beneficiaries Enrolled in Medicare Fee-for-Service (Duals-1 FFS) or Medicare-Medicaid Plans (Duals-1 MMP)

New Measure

Brief Description of Measure: These two measures capture any inpatient or observation stay ("hospitalization") for ambulatory care sensitive conditions (ACSCs) for dually eligible (for both Medicare and Medicaid) beneficiaries 18 years of age and older.

The Duals-1 FFS measure evaluates the performance of all 50 states and the District of Colombia.

The Duals-1 MMP measure evaluates the performance of MMPs. Currently there are 46 MMPs in 9 states.

Both measures report observed and risk-adjusted rates of hospitalizations for ACSCs per 1,000 beneficiaries for three populations ("strata"):

1. Community-dwelling home- and community-based services (HCBS) users;
2. Community-dwelling non-HCBS users (referred to a non-HCBS); and
3. Non-community dwelling population (referred to as Institutionalized).

Both measures are composite measures. Specifically, each is reported as two rates, Acute and Chronic, and as a Total rate, which is a composite of the two. Thus, for each of the three strata, the two measures report three observed rates and three risk adjusted rates:

1. Acute ACSC;
2. Chronic ACSC; and
3. Total (acute and chronic) ACSC

These measures are intended for use in public reporting and quality improvement. These measures can help states and MMPs understand the quality of outpatient care, including home- and community-based services, provided to dually eligible beneficiaries for acute conditions, chronic conditions, and both together. Both measures can assess the quality of a breadth of outpatient services by providers that may not be linked to a single accountable healthcare facility.

One of these measures, the Duals-1 FFS, is currently endorsed (NQF #3449) and due for endorsement maintenance, however we are submitting it together with Duals-1 MMP (a new measure) as new measures. The originally endorsed measure has been substantially modified and aligned with the new Duals-1 MMP measure.

Numerator Statement: These measures report the observed rate and risk adjusted ratio of observed to expected hospital inpatient and observation stays ("hospitalizations") for ACSCs per 1,000 dually eligible beneficiaries 18 years of age and older. We produce measure scores for three admission type strata:

1. Acute: Number of hospitalizations in the measurement year for bacterial pneumonia, urinary tract infection, cellulitis, and pressure ulcers.
2. Chronic: Number of hospitalizations in the measurement year for acute bronchitis, diabetes short-term complications, diabetes long-term complications, uncontrolled diabetes, low-

extremity amputation, chronic obstructive pulmonary disease (COPD), asthma, hypertension, or heart failure.
3. Total: Number of hospitalizations for either Acute or Chronic conditions.

Denominator Statement: Duals-1 FFS: Dually eligible adults age 18 years and older within each state

Duals-1 MMP: Dually eligible adults age 18 years and older enrolled in each Medicare-Medicaid Plan

Denominator Exclusions: Duals-1 FFS: Exclude beneficiaries receiving hospice care at the start of the measurement period, and those that reside in a US territory or non-US country.

Duals-1 MMP: Exclude beneficiaries receiving hospice care at the start of the measurement period.

Measure Type: Composite

Data Source: Claims, Enrollment Data

Level of Analysis: Health Plan, Population: Regional and State

Risk-Adjusted: Statistical risk model (37 factors for Duals-1 FFS and 22 factors for Duals-1 MMP), Stratification by HCBS, non-HCBS, and Institutionalized

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-5; M-5; L-0; I-0

Specifications:

- Measure specifications are clear and precise.

Reliability Testing:

- Reliability testing was conducted at the accountable-entity level:
    - For measure score reliability, the developer calculated signal-to-noise reliability scores for each "strata" of the measure and the total measure. The developer defines strata as the following:
        1. Community-dwelling HCBS users
        2. Community-dwelling non-HCBS users (referred to a non-HCBS)
        3. Non-community dwelling population (referred to as Institutionalized)
    - The median reliability for all strata, for both acute and chronic outcomes, as well as the total, for both measures is above 0.87. Of note, the range of reliability across the Duals-1 MMP was lower for the HCBS and institutionalized strata than Duals-1 FFS.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-1; M-8; L-1; I-0

Validity Testing

- Validity testing was conducted at the accountable-entity level:
    - The developer conducted a systematic assessment of face validity:
        - The independent workgroup of seven members' responses to the question: "The measure can be used to distinguish good from poor quality."

- Seven out of seven, or 100 percent, agreed that both measures could be used to distinguish quality. Of the workgroup members, 5/7, or 71 percent, moderately or strongly agreed that both measures could be used to distinguish good from poor quality.

- The developer also conducted an empirical analysis of the internal validity of the measure:
  - The measure includes two component measures: the Duals-1 FFS measure and Duals-1 Medicare/Medicaid Plan measures. Both measures report observed and risk-adjusted rates of hospitalizations for ACSCs per 1,000 beneficiaries for three populations ("strata").
  - Both measures are composite measures. Specifically, each is reported as two rates, acute and chronic, and as a total rate, which is a composite of the two.
  - Thus, for each of the three strata, the two measures report three observed rates and three risk-adjusted rates.
  - To assess internal validity, the developer calculated Spearman's rank order correlation between the acute and chronic components within each measure and each of the strata.
  - The developer hypothesized that states that perform well on one rate should perform well on the other rates within the measure. All measure rates represent an underlying quality construct of a potentially avoidable hospitalization.
  - Duals-1 FFS
    - The developer found that the state-level measure assessment found moderate to strong correlations of acute and chronic components across states for all strata. The developer further stated that the lowest correlation was for the non-HCBS cohort, with a correlation of 0.29, which is comparable to the conventionally "moderate" correlation of 0.30.
  - Duals-1 MMP
    - The developer found that the state-level measure assessment found one statistically significant correlation, with a correlation of 0.66, for the non-HCBS strata. The other two strata did not have a statistically significant correlation.
- The SMP should consider the methods for empirical testing to confirm the measure is analyzing agreement with another authoritative source of the same information, correlation of measure scores with another valid indicator of quality for the specific topic, or relationship to a conceptually related topic. (Measure evaluation guidebook 2b1.)

Exclusions

- The measure excludes observations for which there is no continuous enrollment for 18 months (measurement and risk adjustment period).
  - Enrolled in FFS with full Medicaid benefits – no continuous enrollment for 18 months 2,151,076 (37.9 percent)
  - Enrolled in MMP – no continuous enrollment for 18 months 208,284 (45.6 percent)

- The developer notes that the exclusions applied were chosen to produce valid, reliable, and fair scores. The developer further emphasizes that continuous enrollment is required to capture both risk factors and outcomes.

Risk Adjustment

- The measure uses a statistical methodology employing a zero inflated negative binomial (ZINB) model to predict the number of ACSC discharges during the measurement period.
- Separate models are estimated for acute and chronic composite outcomes and for each subgroup (i.e., institutional, HCBS, and non-HCBS), resulting in six risk adjustment models for each measure (Duals FFS and Duals MMP).
- The Duals-1 FFS includes 37 risk factors, and the Duals-1 MMP includes 22 risk factors.
- Rurality was retained in all risk adjustment models.
- The developer presents six individual sets of calibration statistics, C-statistics, and calibration plots.
    - Duals-1 FFS
        - The C-statistics for all models indicated discrimination between patients with no events and those with one or more events (ranging from 0.685–0.848).
    - Duals-1 MMP
        - The C-statistics for all models indicated discrimination between patients with no events and those with one or more events (ranging from 0.721–0.809).

Meaningful Differences

- The developer identified a substantial number of outliers across strata and outcomes.
- The developer notes that there were more outliers for the FFS measure than the MMP measure.
- For the FFS measure, between 31–41 percent of states had confidence intervals for the Total rate ratio lying above one, indicating a worse performance than average, and between 25–43 percent of states had confidence intervals below one, indicating a better performance than average. There were similar patterns for acute and chronic composite scores and across all three strata.
- The developer notes that the MMP measure had fewer outliers, reflecting the smaller denominators, but there were still meaningful numbers of high and low performers across all strata and outcomes. For the total composite, 35–41 percent of MMPs had total composite rate ratios that were entirely below one, indicating a higher-than-expected performance. The fewest outliers were for institutional strata, for which there was/were one acute, five chronic, and four total low performers.

Missing Data

- The developer notes that there were no missing data.

Comparability

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

Composite

Preliminary ratings for composite construct: The SMP Passed on composite with a score of: H-2; M-6; L-1; I-1

Empirical analysis to support composite construction

- The developer used the internal consistency of the outcomes for each component and for the composite to support the overall quality construct for the measure.
- The developer calculated Cronbach's alpha across all strata and at the state/MMP level for all outcomes (acute, chronic, and total outcomes).
- The developer provided the following results: Duals-1 FFS: All alphas (acute, chronic, total) were over 0.85; Duals-1 MMP: All alphas (acute, chronic, total) were over 0.85.

SMP Concerns

- Several SMP members raised concerns that the empirical validity testing presented examined the correlation of the two measure components rather than a comparison to another standard metric, which is preferrable. However, all reviewers acknowledged that the face validity testing was adequate and acceptable for a new measure.
- One SMP member noted that regarding meaningful differences, the developer did not appear to present a range or distribution of rates across states or MMRs and that the presented data (statistically better/worse analysis with small sample sizes) is not meaningful or sufficient.
- Most SMP members did not have concerns with missing data or exclusions. However, two reviewers did note that a large proportion of patients were excluded from the sample because they did not have 18 months of Medicare/Medicaid eligibility. They noted that this large number of exclusions should be discussed by the Standing Committee.
- Most SMP members had no concerns regarding the risk adjustment model and approach to testing. However, some noted that the factors included and excluded in the model should be carefully reviewed by the relevant Standing Committee as many factors were excluded.
- A small number of SMP members raised concerns with the composite construct. One member noted that the developer did not provide sufficient justification for combining the measures, stating that acute and chronic events are different enough to merit separate measures. Another member noted that it is difficult to know whether the rate for the composite is heavily weighted and influenced by the relative rate or relative variance of one component or another because the developer did not provide raw or adjusted rates or a distribution of rates for the two components, nor did they provide the rates of admission for each ambulatory care sensitive condition.

NQF #2651 CAHPS® Hospice Survey, Version 9.0

Maintenance Measure

Brief Description of Measure: The measures submitted here are derived from the CAHPS® Hospice Survey, Version 9.0, a 39-item standardized questionnaire and data collection methodology. The survey is intended to measure the care experiences of hospice decedents and their primary caregivers. Survey respondents are the primary informal caregivers (i.e., family members or friends) of patients who died while receiving hospice care.

The proposed measures include the following six multi-item measures:

- Hospice Team Communication
- Care Preferences
- Getting Timely Care
- Treating Family Member with Respect
- Getting Emotional and Religious Support

- Getting Help for Symptoms

In addition, there are three single-item measures:

- Getting Hospice Training
- Rating of the Hospice
- Willingness to Recommend the Hospice

Following is a list of the survey items included in each measure.

Hospice Team Communication (5 items)

- How often did the hospice team keep you informed about when they would arrive to care for your family member?
- How often did the hospice team explain things in a way that was easy to understand?
- How often did the hospice team listen carefully to you when you talked with them about problems with your family member's hospice care?
- How often did the hospice team keep you informed about your family member's condition?
- While your family member was in hospice care, how often did the hospice team listen carefully to you?

Care Preferences (2 items)

- Did the hospice team make an effort to listen to the things that mattered most to you or your family member?
- Did the hospice team provide care that respected your family member's wishes?

Getting Timely Care (2 items)

- When you or your family member asked for help from the hospice team, how often did you get help as soon as you needed it?
- How often did you get the help you needed from the hospice team during evenings, weekends, or holidays?

Treating Family Member with Respect (2 items)

- How often did the hospice team treat your family member with dignity and respect?
- How often did you feel that the hospice team really cared about your family member?

Getting Emotional and Religious Support (3 items)

- While your family member was in hospice care, how much emotional support did you get from the hospice team?
- In the weeks after your family member died, how much emotional support did you get from the hospice team?
- Support for religious or spiritual beliefs includes talking, praying, quiet time, or other ways of meeting your religious or spiritual needs. While your family member was in hospice care, how much support for your religious and spiritual beliefs did you get from the hospice team?

Getting Help for Symptoms (4 items)

- Did your family member get as much help with pain as he or she needed?
- How often did your family member get the help he or she needed for trouble breathing?
- How often did your family member get the help he or she needed for trouble with constipation?
- How often did your family member get the help he or she needed from the hospice team for feelings of anxiety or sadness?

Getting Hospice Care Training (1 item)

- Hospice teams may teach you how to care for family members who need pain medicine, have trouble breathing, are restless or agitated, or have other care needs. Did the hospice team teach you how to care for your family member?

Rating of Hospice Care (1 item)

- Using any number from 0 to 10, where 0 is the worst hospice care possible and 10 is the best hospice care possible, what number would you use to rate your family member's hospice care?

Willingness to Recommend Hospice (1 item)

- Would you recommend this hospice to your friends and family?

A complete list of proposed CAHPS Hospice Survey measures, including response options for each item, is available in Appendix B.

Numerator Statement: CMS calculates CAHPS Hospice Survey measure scores using top-, middle- and bottom- box scoring. The top-box score refers to the percentage of caregiver respondents that give the most positive response(s). The bottom box score refers to the percentage of caregiver respondents that give the least positive response(s). The middle box is the proportion remaining after the top and bottom boxes have been calculated; see below for details. Details regarding the definition of most and least positive response(s) are noted in Section SP.14 below.

Denominator Statement: In national implementation and public reporting, CAHPS® Hospice Survey measure scores are calculated only for hospices that had at least 30 completed questionnaires over the most recent eight quarters of data collection.

The target population for the survey are the adult primary caregivers of hospice decedents. Respondent eligibility and exclusions are defined in detail in the sections that follow. A survey is defined as completed when at least 50 percent of the questions applicable to all decedents/caregivers are answered. The survey uses screener questions to identify respondents eligible to respond to subsequent items. Therefore, denominators vary by survey item (and corresponding multi-item measures, if applicable) according to the eligibility of respondents for each item. In addition, for the Getting Hospice Care Training measure, scores are calculated only among those respondents who indicate that their family member received hospice care at home or in an assisted living facility.

Denominator Exclusions: The exclusions noted in here are those who are ineligible to participate in the survey. The one exception is caregivers who report on the survey that they "never" oversaw or took part in the decedent's care; these respondents are instructed to complete the "About You" and "About Your Family Member" sections of the survey only.

Cases are excluded from the survey target population if:

- The hospice patient is still alive

- The decedent's age at death was less than 18
- The decedent died within 48 hours of his/her last admission to hospice care
- The decedent had no caregiver of record
- The decedent had a caregiver of record, but the caregiver does not have a U.S. or U.S. Territory home address
- The decedent had no caregiver other than a nonfamilial legal guardian
- The decedent or caregiver requested that they not be contacted (i.e., by signing a no publicity request while under the care of hospice or otherwise directly requesting not to be contacted)
- The caregiver is institutionalized, has mental/physical incapacity, has a language barrier, or is deceased
- The caregiver reports on the survey that he or she "never" oversaw or took part in decedent's hospice care

Measure Type: Outcome: PRO-PM

Data Source: Instrument-Based Data

Level of Analysis: Facility

Risk-Adjusted: Statistical model with nine risk factors

Sampling Allowed: Yes

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-6; M-3; L-2; I-0

Specifications:

- Measure specifications are clear and precise.
- Measure specifications have been updated since the 2019–2020 NQF maintenance endorsement.
  - Specifically, the survey instrument was revised based on feedback from a developer-convened TEP and public comments during the 2019–2020 maintenance cycle to shorten and simplify the instrument as well as to add a new two-item Care Preferences measure.
- Measure specifications include the specific instrument (e.g., patient-reported outcome measure [PROM]); standard methods, modes, and languages of administration; whether proxy responses are allowed; standard sampling procedures; and handling of missing data.

Reliability Testing:

- Updated testing included 56 hospices participating in the 2021 CAHPS Hospice Survey Mode Experiment, with 5,731 total responses.
- Reliability testing was conducted at the patient/encounter level:
  - Multi-item measure reliability was assessed using Cronbach's alpha (internal consistency). Cronbach's alpha was >=0.70 for five of six multi-item measures; it was 0.62 for one multi-item measure (Getting Timely Care). Cronbach's alpha when an item was deleted decreased for all but one item.

- Multi-item measure reliability was also assessed using the person-level Pearson item-total correlation (relation of each item to all other items). Item-total Pearson correlation ranged from 0.45–0.71.
- Reliability testing was conducted at the accountable-entity level:
  - Inter-unit (hospice-level) reliability was calculated at the mean sample sizes, using intraclass correlation coefficients (ICCs) calculated from the case mix-adjusted 0–100 top-box scores, applying the Spearman Brown prediction formula. Hospice-level reliability at the average number of respondents ranged from 0.70–0.84 on six multi-item measures and 0.70–0.87 on three single-item measures.
  - The developer also cites published research assessing the stability of responses to items that assess the overall quality of care and willingness to recommend with the agreement of 86 percent or higher for overall quality and 90 percent or higher for the willingness to recommend. Kappa statistics ranged from 0.58 for the willingness to recommend to 0.70 for overall quality in repeated measures.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-1; M-6; L-2; I-2

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - Exploratory and confirmatory factor analyses (CFA) of newly tested items and items in multi-item measures were conducted using weighted least squares means and variance adjusted (WLSMV) estimation. The assessed overall model fit for the six-factor model using the comparative fit index was 0.997, the root mean square error of approximation was 0.014, and the weighted root mean square residual was 1.068 in CFA. The factor loadings were above 0.70. The overall fit chi-squared was 120, which equaled 252.83, $p<0.001$.
  - Construct validity was assessed using Pearson correlations between six multi-item and one single-item measure top-box scores and with two single-item global measures top-box scores. The Pearson correlations ranged from 0.40–0.61 across the measures.
  - Discriminant validity was assessed using Pearson correlations among multi-item measures to evaluate the extent to which they measure different constructs. The Pearson correlations ranged from 0.33–0.64.
- Validity testing was conducted at the accountable-entity level:
  - The developer noted that it used individual-level data for updated testing "as estimates of hospice-level associations would be unbiased but imprecise if calculated among the 56 hospices participating in the 2021 mode experiment."
  - Prior testing (the 2019 submission) included both individual- and hospice-level results. The hospice-level Pearson correlations between measures and global rating items ranged from 0.63–0.84. The hospice-level Pearson correlations among multi-item measures ranged from 0.42–0.84.

Exclusions

- Decedents or caregivers who otherwise meet the inclusion criteria are excluded if they have a "no publicity" status. No statistical testing was conducted given the nature of this exclusion.

Risk Adjustment

- The model was developed during the prior maintenance review. It was not retested or updated for this submission.
- There is a risk model with nine risk factors (i.e., response percentile, decedent age, payer, primary diagnosis, length of final episode of hospice care, respondent age, respondent education, relationship of decedent to caregiver, and language).
  - Case-mix coefficients from a linear regression are used to generate case-mix adjustments for each survey question.
  - Publicly reported hospice survey measure scores are adjusted to the overall national mean of case-mix variables across all reporting hospices.
- Published literature and data analyses were used to develop the conceptual model and select the risk adjustment approach.
  - The developer identified characteristics as candidates for adjustment if they were present in the response data and not in the hospice's control. For each adjuster, they examined variation among hospices using ICC, bivariate and multivariate association with selected survey outcomes, and the impact on adjustment and parameterization of adjustors.
  - The following social risk factors were considered: decedent education, primary payer for hospice care, caregiver respondent education, and caregiver respondent language. Primary payer and language are included in the risk adjustment model. The two education variables were associated with the outcomes but moderately correlated with each other; therefore, caregiver education was retained while decedent education was not.
- No discrimination or calibration statistics were provided.

Meaningful Differences

- The developer calculated number and percentages of hospices significantly above or below the mode experiment hospice average for each measure; scores were adjusted for mode and case mix. Between 13 and 26 percent of hospices were statistically different (above or below) the mode experiment hospice average.

Missing Data

- Survey response rates ranged from 31–45 percent across modes. Item-level missing data due to inappropriate skips ranged from 0.5–5.0 percent.
- The developer cites prior research that indicates that nonresponse weighting to account for potential bias is not needed after case-mix adjustment.

Comparability

- Linear regression was used to evaluate the effects of different survey modes on survey outcomes. The model included case-mix adjusters, hospice indicators, and the month of death.
- There were significant effects of survey mode on several survey outcomes. Consequently, the survey scores should be adjusted for the mode of administration.

SMP Concerns

- One member noted that at the hospice level the submission refers to above or below the average of experiment's participating hospices and it is whether the developers mean top box scores or actual mean scores. The same member noted concerns with case-mix adjustment particularly around language and mode of administration, sampling due to possible bias introduced from the poor response rate, and consistency for vendors with multiple hospices.

- Regarding patient/encounter level reliability, one member stated that the submission notes differences in rating within each domain based on characteristics of decedent, respondent, and mode of survey administration. The member continued that although risk adjustment attempts to correct for these, the methodology ignores the possibility that there are systematic differences in performance associated with decedent characteristics that should not be adjusted away.
- Regarding accountable-entity level reliability, one member noted that per their comment regarding the risk adjustment methodology, the differences in performance across different patients has not been explored. A second member noted that the hospice level reliability (0.03) was poor.
- Regarding the patient/encounter level validity testing, one SMP member noted that within survey validity measures create over endogeneity of results and therefore it is common to use more general/global measures of the quality construct to validate construct-specific measures. Additionally, it was noted that the scale quality is high enough but there is evidence that hospices may differ along different dimensions of care and because of this a factor analysis could have been useful to present.
- Regarding the accountable-entity level testing and the risk adjustment model, some reviewers noted that the testing was not updated even though the developers updated the survey. Another reviewer noted that the developer did not provide entity level results, rather they relied on respondent-level correlations between measures.
- Regarding missing data, some SMP members noted that there is concern with large non-response, however the developer addresses these concerns in the risk adjustment model. Further, an SMP member noted that having characteristics of non-responders would have been helpful to assess whether certain caregivers known to have poorer or better ratings of hospices were more frequently non-responsive.
- The developer provided responses to these concerns, which are available in Appendix B.

NQF #3726 Serious Illness Survey for Home-Based Programs

New Measure

Brief Description of Measure: The proposed measures are derived from the Serious Illness Survey for Home-Based Programs, a 36-item questionnaire designed to measure the care experiences of patients receiving care from home-based serious illness programs. Home-based serious illness programs provide care for seriously ill patients at their private residences (i.e., in their homes or assisted living facilities, not in institutions like skilled nursing facilities). Programs are staffed by interdisciplinary teams that provide support for palliation of symptoms, assist with coordination of care, answer questions after-hours, provide medication management, and assist with advance care planning (Cohn et al., 2017). Teams consist of clinicians (e.g. physicians, nurse practitioners) that oversee care, as well as clinical and supportive staff that make home visits (e.g. registered nurses, social workers, CNAs). Programs serve patients with a life expectancy that ranges from 1-5 years and have enrollment criteria based on diagnosis, symptom burden, functional status, and/or prior health care utilization.

The five proposed multi-item measures are:

1. Communication
2. Care Coordination
3. Help for Symptoms
4. Planning for Care

5. Support for Family and Friends

The two proposed single-item measures are:

1. Overall Rating of the Program
2. Willingness to Recommend the Program

Appendix A presents the survey items included in each measure, including response options for each item. Measure scores are "top-box" scores that reflect the percent of respondents who select the most positive response category(ies) in response to the survey item(s) within the measure.

Citation:

Cohn J, Corrigan J, Lynn J, Meier D, et al. Community-Based Models of Care Delivery for People with Serious Illness. National Academy of Medicine Discussion Paper. Available at https://nam.edu/wp-content/uploads/2017/04/Community-Based-Models-of-Care-Delivery-for-People-with-Serious-Illness.pdf.

Numerator Statement: Measure scores are "top-box" scores that reflect the percent of respondents who select the most positive response category(ies) in response to the survey item(s) within the measure. Therefore, the numerator is the number of respondents who select the most positive response category(ies) in response to the survey items within the measure.

Denominator Statement: Survey respondents are patients receiving care from home-based serious illness programs. Survey eligibility criteria and exclusions are detailed below in sections sp.16 – sp.18. Screener questions and tailored non-applicable response options (e.g., I did not want help for my pain) are used to identify respondents who are and are not eligible to respond to survey items included in evaluative measures. Therefore, denominators vary by survey item (and corresponding multi-item measures, if applicable) according to the eligibility of respondents for each item.

Denominator Exclusions: The Serious Illness Survey for Home-Based Programs is designed for administration to adult patients who are currently enrolled in home-based serious illness programs. Patients are excluded from the survey sample if they:

- Are under age 18
- Receive care from a serious illness program in a setting OTHER than home or an assisted living facility (e.g., in a nursing home or other long-term care facility)
- Are known to have been discharged to hospice
- Are known to have died
- Have been enrolled in the serious illness program for less than six weeks as of the date of survey sampling

In keeping with the Medicare CAHPS Survey (https://www.cms.gov/files/document/ma-pdp-cahps-qapts-v11-complete-manual.pdf), a survey is considered partially completed if there are responses to at least one measure and for less than 50 percent of survey items that are applicable to all. A survey is considered completed if there are responses to at least one measure and for 50 percent or more of the survey items that are applicable to all. Final analytic datasets include all completed and partially completed surveys.

There are no explicit exclusions based on language; the survey is available in English and Spanish.

Measure Type: Outcome: PRO-PM

Data Source: Instrument-Based Data

Level of Analysis: Other: Home-based Serious Illness Program

Risk-Adjusted: Statistical risk model with eight factors

Sampling Allowed: Yes

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-4; M-4; L-2; I-1

Specifications:

- Measure specifications are clear and precise.
- Measure specifications for the instrument-based measure also include the specific instrument (e.g., PROM[s]); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; and the calculation of response rates to be reported with the performance measure results.

Reliability Testing:

- Testing data are from 32 Serious Illness Programs with a total of 2,263 respondents. Eligible patients were randomly assigned one of two modes of administration: mail-only or telephone-only.
- Reliability testing was conducted at the patient/encounter level:
  - Cronbach's alpha was used to assess the internal consistency of multi-item measures. Cronbach's alpha was greater than or equal to 0.70 for four of five multi-item measures. For the fifth measure, it was 0.69. Cronbach's alpha with item deletion was lower.
  - Pearson item-total correlation was also calculated. The Pearson correlation ranged from 0.44–0.69.
- Reliability testing was conducted at the accountable-entity level:
  - Program level reliability was calculated using ICCs of case-mix and survey-mode adjusted top-box scores for programs with 10 or more respondents (28 of 32 programs). Predicted program-level reliability was calculated using the Spearman–Brown formula at 100 respondents. Program-level reliability at 100 measure respondents ranged from 0.67–0.80. Values were greater than 0.70 for all but one measure (single-item global measure of Rating of Program).

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-3; M-6; L-2; I-0

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - Confirmatory factor analyses of 18 survey items were identified by the TEP as most important using WLSMV. The assessed overall model fit for the six-factor model using comparative fit index was 0.992. The root mean square error of approximation was 0.023. The weighted root mean square residual was 1.463. The factor loadings were above 0.70. The overall fit chi-squared was 125, which equaled 269.45.
- Validity testing was conducted at the accountable-entity level:

- Construct validity was assessed by examining the associations between each multi-item measure top-box score with two single-item global measures top-box scores. The developer estimated multivariate linear regression models with the global measures as dependent variables. Models were adjusted for case-mix and survey mode and estimated with weighted least square mean and variance adjustment. Standardized regression coefficients ranged from 0.44–0.57 across the measures.
- Discriminant validity was assessed using correlations among multi-items computed as the average of top-box scored items. The correlations ranged from 0.39–0.62.

Exclusions

- Exclusions are included to ensure that only those who receive care from home-based serious illness programs and have sufficient experience with the program are in the sample. Exclusions are not based on statistical testing.

Risk Adjustment

- There is a risk model with eight risk factors (age, education, primary diagnosis, proxy response, self-reported functional status, self-reported physician health, self-reported mental health, and response percentile).
  - A fit linear model for each survey item includes each case mix adjustor, survey mode, and program fixed effects.
  - The model is used to generate adjusted scores for each program.
- Published literature and data analyses were used to develop the conceptual model and select the risk adjustment approach.
  - The developer identified characteristics as candidates for adjustment exogeneous to the care provided by the program.
  - Linear regression models were used to estimate the effect of potential case-mix adjustors on survey measure scores. The impact of each adjustor on program-level scores was also evaluated. The criteria for inclusion in the final model were statistically significant and associated with at least one quality measure score at the 0.01 level and $1 - r^2$ of at least 1 percent for at least one quality measure score.
  - Testing also incorporated feedback from the TEP for face validity.
  - The following social risk factors were considered: language, education, and payer. Payer data were missing for about 20 percent of respondents and were removed from consideration. The education met the testing criteria for model inclusion; the language did not.
- **Discrimination.** The R-squared values ranged from 0.06–0.12 (for each measure). The developer also compared program scores with and without adjustment using Kendall's tau. Kendall's tau ranged from 0.76–0.93. The percentage of program pairs that would switch rankings ranged from 4–12 percent.
- **Calibration.** The developer calculated the correlation between observed and model-predicted values for each measure. The correlation values ranged from 0.25–0.35.
- The developer notes that it would not expect case-mix adjustment to predict a great deal of variation in response (versus what might be expected to be seen for clinical outcomes).

Meaningful Differences

- The developer calculated the number and percentage of programs significantly above or below the field test program average for each measure. Top-box scores were adjusted for mode and case mix. Between 18–29 percent of programs scored either above or below the field test program average.

Missing Data

- Survey response rates were 30.4 percent for mail-only and 42.5 percent for mail-telephone. Item-level missing data due to inappropriate skips ranged from 0.9–6.4 percent.
- The developer cites prior research that indicates that nonresponse weighting to account for potential bias is not needed after case-mix adjustment.

Comparability

- Linear regression was used to evaluate the effects of mode on measure scores.
- The statistically significant effects of mode on measures scores were not detected. However, the developer recommends adjusting for mode when calculating scores.

SMP Concerns

- Some SMP members expressed that the accountable-entity level reliability testing results reported were low. Additionally, one member expressed concern that it is unclear if the measure should be used when there are fewer than 10 respondents.
- A number of SMP members noted concerns that in the risk adjustment assessment it is not clear how much scores changed rankings and whether adjustment is needed. One member noted concerns that variance explained by case-mix adjustment is less than 12 percent. Another member noted that the model cannot adequately explain the variation in responses and entity scores since the result is similar with and without risk adjustment.
- The developer provided responses to these concerns, which are available in Appendix B.

NQF #3722 Home Dialysis Rate

New Measure

Brief Description of Measure: Percent of all dialysis patient-months in the measurement year in which the patient was dialyzing via a home dialysis modality.

Numerator Statement: Patient-months from the denominator in which the patient was dialyzing via a home modality (peritoneal dialysis and/or home hemodialysis) as of the final dialysis treatment of the given measurement month.

References:

1. A patent-month construct is used to account for patients' potentially varying time contributions to both the numerator and denominator.

Denominator Statement: All dialysis patient-months (in-center and/or home) attributed to the dialysis facility (or aggregate HRR unit)[2] during the measurement year.

References:

1. A patent-month construct is used to account for patients' potentially varying time contributions to both the numerator and denominator.

2. In recognition of the structure of the dialysis market, if a company (e.g., dialysis organization) owns multiple facilities in a given Hospital Referral Region (HRR), it would report an aggregated score for all facilities located within the HRR owned wholly or in part by the company.

Denominator Exclusions: The following exclusions are applied to the denominator:

1. Patient-months in which the patient was admitted to the facility to which they are attributed for less than 30 days as of the final day of the measurement month
2. Patient-months in which the patient is receiving dialysis for AKI only at any time in the measurement month
3. Patient-months in which the patient is enrolled in hospice at any time in the measurement month
4. Patient-months in which the patient is residing in a nursing home or other LTCF at any time in the measurement month
5. Patient-months in which the patient was discharged from the facility secondary to transplant, death, discontinuation of dialysis, and/or recovery of function at any time in the measurement month

Measure Type: Process

Data Source: Electronic Health Data, Electronic Health Records

Level of Analysis: Facility, Other, Please Explain: To account for home dialysis–only facilities within a Hospital Referral Region (HRR), particularly if a parent company sends its home dialysis patients to such a provider, the measure allows for aggregation of facilities owned by the same company within a given HRR. Specifically, a subsidiary facility's Home Dialysis Rate is aggregated to the facility's aggregation group, which includes all dialysis facilities owned in whole or in part by the same legal entity ("Parent Organization") located in the HRR in which the facility is located.

Risk-Adjusted: Stratification by five risk factor groups

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-5; M-3; L-1; I-2

Specifications:

- This measure was previously submitted to the SMP under NQF #3679 as a clinical intermediate outcome measure. The developer has resubmitted it as a process measure under NQF #3722.
- To account for previous feedback, the developer updated the level of analysis from facility to facility and HRR. Additionally, the developer updated their reliability testing (see below) to account for the nonindependence of patient months.
- Measure specifications are clear and precise.

Reliability Testing:

- Reliability testing was conducted at the accountable-entity level:
  - Reliability testing was conducted at the facility and HRR levels using signal-to-noise analysis: the beta-binomial model. HRR-level analysis was completed for facilities with common ownership aggregated within the HRR.

- The mean reliability at the facility level was 0.999 (median =1) and the HRR-level reliability was 0.994 (median 0.997). The developer further examined the HRR level by presenting the data at two parent dialysis organizations (DOs). DO 1 reliability was 0.994 (median 0.998), and DO 2 reliability was 0.997 (median 0.999).
- The developer also calculated mean reliability based on the number of patients per facility (<25, 25–49, 50–79, 80–119, and 120+). Mean reliability was 0.996 for the smallest facilities (<25 patients).
- In order to address the potential for nonindependence of patient months biasing the reliability estimates, the developer calculated monthly reliability estimates. The minimum mean reliability in a month was 0.986 at the facility level and 0.919 at the HRR level.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-1; M-6; L-3; I-1

Validity Testing

- Validity testing was conducted at the accountable-entity level:
  - Validity testing was conducted using face validity with a panel of nine members (five healthcare providers, two dialysis facilities, and three manufacturer groups).
  - Eight of the nine members agreed that the measure score is likely or highly likely to provide an accurate reflection of quality and that the measure would effectively distinguish real differences in performance between providers.
  - Eight of the nine members agreed that the measure scores for the paired set (NQF #3722 and NQF #3725) will provide an accurate reflection of quality and that the paired set will effectively distinguish real differences in performance between providers.
  - The one dissenting member noted concerns about the minimal patient exclusion criteria and that this would make the measure more of a reflection of the provider's patient population and not their performance.

Exclusions

- The following exclusions are applied to the denominator: patient months with less than 30 overall days in a facility (4.2 percent), patients months with acute kidney injury (AKI) (2.0 percent), patient months with hospice (0.0 percent), patient months in nursing home or other LTCF (2.8 percent), and patient discharge secondary to transplant (0.2 percent), death (1.2 percent), discontinuation of dialysis (0.2 percent), and/or recovery of renal function in the month (0.2 percent). After accounting for overlap in exclusions, a total of 9.5 percent unique patient months were excluded from the denominator. Mean facility level performance before exclusions was 13.28 percent and with them applied was 14.49 percent. HRR aggregated facility level performance was provided for two DOs.
- Mean performance before exclusions was 15.94 percent (DO 1) and 14.32 percent (DO 2); with exclusions applied, it was 17.26 percent (DO 1) and 16.37 percent (DO 2). The developer believes that these exclusions are clinically warranted to minimize the capture of patients for whom home dialysis is not suitable, desirable, or relevant.

Risk Adjustment

- The developer stratified the measure by age, gender, race/ethnicity, and dual-eligible status. They also explored markers of functional risk and clinical variables for stratification, but they were not included due to data availability.

- Stratified analyses at both the facility and HRR levels demonstrate a clear trend by age (as age increases, the percent on home dialysis falls), differences by race (the percent for White is higher than for Black patients but less than for "Other" race), and that the percent on home dialysis is less among dual-eligible than non-dual-eligible patients.

Meaningful Differences

- Over half of the facilities have zero patient months with home dialysis, and the 75th percentile is 20.02 percent of patient months with home dialysis. At the HRR level, the 25th percentile performance is 12.52 percent, the median is 16 percent, and the 75th percentile is 19.51 percent.
- To demonstrate the statistical significance of the spread at the facility level, the developer analyzed 3,071 facilities with a non-zero performance score. The overall weighted mean performance score was 24.5 percent with the facility size as the weight. The developer noted this as the national norm. Facilities with a score between 0.05 percent and 6.47 percent all had 95 percent CIs below the norm (below the 20th decile). Facilities with a score from greater than 95.3–100 percent all had 95 percent CIs above the norm (90th decile and above). Facilities with a score between greater than 36.8–100 percent had 95 percent CIs above the norm (80th decile and above).

Missing Data

- The developer notes that while they believe their observed percent of patient-months excluded secondary to hospice enrollment is an underestimate, they believe those same patients are captured in other exclusions. The developer also believes their observed percent of patient-months excluded due to nursing/LTCF residence is an underestimate. However, they note that if they were to use the highest exclusion rate reported, there is only a difference of 0.3 percent in the overall facility-level score.
- When patient months were excluded from the denominator due to missing values in the stratification variables (i.e., age, sex, race, ethnicity, and dual-eligibility status), the mean facility level performance dropped by 0.09 percent at the facility level and 0.11 percent at the HRR level after excluding missing values.

Comparability

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

SMP Concerns

- Some SMP members raised concerns with the reliability testing and the details of the facility-level calculation noting that the results at the facility level are unusually high. One member noted that, given the variance in rates of use of home dialysis across units, the reliability scores would be expected to be lower.
- There was disagreement about the measure's ability to distinguish meaningful differences and whether the testing adequately addressed facilities with zero months of home dialysis. Some voiced specifically that the testing data does not have a normal distribution of performance at the facility level since over half of facilities have zero months with home dialysis. The measure therefore cannot differentiate adequately. However, others did not have this concern and found the demonstration of meaningful difference adequate because testing was performed on facilities with non-zero performance scores.

- Some SMP members raised concerns with the risk stratification approach used by the developers, noting a lack of detail regarding the methodology for stratification.

Subgroup 2

NQF #2958 Informed, Patient-Centered (IPC) Hip and Knee Replacement Surgery

Maintenance Measure

Brief Description of Measure: The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60 percent or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision.

The target population is adult patients who had a primary hip or knee replacement surgery for treatment of hip or knee osteoarthritis.

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

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

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

Measure Type: Outcome: PRO-PM

Data Source: Instrument-Based Data

Level of Analysis: Clinician: Group/Practice

Not Risk-Adjusted

Sampling Allowed: Yes

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-6; M-2; L-0; I-1

Specifications:

- Measure specifications are clear and precise.
- Measure specifications for the instrument-based measure also include the specific instrument (e.g., PROM[s]); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and the calculation of response rates to be reported with the performance measure results.

Reliability Testing:

- Reliability testing was conducted at the patient/encounter level:

  - During the 2016 submission, the developer conducted test-retest reliability of the knowledge and preference items from the same individuals four to six weeks apart.
  - For the knowledge score, the developer examined the ICC of the knowledge score at time #1 and time #2.
  - For the preference item, the developer examined the kappa between the response at time #1 and response at time #2.
  - The test-retest reliability of the knowledge score was examined in sample #1 with an ICC of 0.81 (95 percent CI ranging from 0.71–0.87). The test-retest reliability of the item assessing preferred treatment had a Kappa of 0.801.
- Reliability testing was conducted at the accountable-entity level:
  - For the current submission, the developer divided data within each practice site into samples with a minimum size of 50. The percentage with IPC within each sample was calculated.
  - The reliability was calculated as variability from site divided by total variability. The developer reported that for four groups (site 1 had 16 samples, site 2 had 26 samples, site 3 had 26 samples, and site 4 had four samples), the reliability was 0.735. In the 2016 submission, the developer found that for 14 groups (site 1 had two samples, site 2 had seven samples, site 3 had two samples, and site 4 had three samples), the reliability was 0.853.
  - The developer noted that the reliability estimate is slightly lower than the prior submission due to the randomization of individuals to groups.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-4; M-4; L-1; I-0

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - For the 2016 submission, the developer reported patient/encounter level validity testing.
  - The developer performed discriminant validity of the knowledge assessment by comparing scores of those who should have higher knowledge (e.g., scores of patients who had used a decision aid versus those who did not). The developer stated that the mean knowledge scores discriminated between patients in a decision aid group with 67 percent (SD of 21.2) compared to 51 percent (SD of 24.9) in the usual care group ($p<0.001$).
  - The developer also examined whether patients who stated a clear preference for surgery rated the importance of relieving pain and improving function higher than those who were unsure or those who stated a preference for nonsurgical treatments. Further, the developer examined whether those who stated a clear preference for surgery rated the importance of avoiding surgery lower than those who were unsure or those who stated a preference for nonsurgical treatments. These hypotheses were tested using analysis of variance (ANOVA) with planned comparisons.
  - The data provide evidence that the measure can discriminate among groups with different levels of knowledge (e.g., those who have viewed a decision aid or not), and the preference item can discriminate among patients who place a different amount of importance on salient goals relating to treatment for osteoarthritis.
- Validity testing was conducted at the accountable-entity level:

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

Exclusions

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

Risk Adjustment

- The measure is not risk-adjusted or stratified.
- The developer states that it does not recommend risk adjustment for this measure. Any patient who has one of these elective surgeries should be able to answer the knowledge questions

correctly and should have a clear preference for the procedure (to meet the standards of informed consent).

Meaningful Differences

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

Missing Data

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

Comparability

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

SMP Concerns

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

NQF #2962 Shared Decision-Making Process

Maintenance Measure

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

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

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

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

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

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

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

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

Measure Type: Outcome: PRO-PM

Data Source: Instrument-Based Data

Level of Analysis: Clinician: Group/Practice

Not Risk-Adjusted

Sampling Allowed: Yes

Reliability

Preliminary ratings for reliability: The SMP Passed on Reliability with a score of: H-0; M-8; L-0; I-2

Specifications:

- Measure specifications have not changed since the last review.
- Measure specifications are clear and precise.
- Measure specifications for the instrument-based measure also include the specific instrument (e.g., PROM[s]); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and the calculation of response rates to be reported with the performance measure results.

Reliability Testing:

- For this current submission, the developer presents reliability testing conducted at the accountable-entity level:

- The developers divided patients from the same site making the same decisions into random groups and correlated their process scores. The developer implemented a minimum sample size of 58, which results in 76 patient groups from 5,294 patient reports. The developers reported an average reliability of 0.69 (95 percent CI = [0.685, 0.69])
- The developers also reported an ICC by dividing the between site variance by the total variance resulting in an ICC of 0.96.

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

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-3; M-4; L-1; I-2

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - In the current submission, the developer provides evidence from three published studies, which depict the relationship of this measure in the predicted direction with other decision-making outcomes (e.g., higher confidence; satisfaction; less regret; and higher rates of informed, patient-centered surgery).
    - For the Valentine et al 2021a paper, an effect size was calculated using a model of inverse variance methods and random effects. The heterogeneity of the

effects was calculated using the DerSimonian–Lair estimator of between-study variance.

- The developer attests that Valentine et al 2021a's results showed that the SDM process scores were related to higher decision confidence (effect size = 0.57, p<0.001); lower decision regret (effect size = -0.34, p<0.001); and higher rates of informed, patient-centered decisions (effect size = 0.18, p=0.03).

- For the Brodney 2019 paper, the developers used generalized linear and logistic regression models with the General Estimating Equations approach to account for the clustering of patients within surgeons. The developer states that models were adjusted for patient characteristics, such as age, gender, education, joint, and baseline quality of life scores.
  - The developer attests that Brodney 2019's results showed that SDM process scores were higher among patients who reported no regret (2.5 [1.2] no regret versus 2.3 [1.2] regret, p<0.001 for hip and knee surgery); higher among patients who reported high satisfaction ([2.3 (1.2) not satisfied vs. 2.5 satisfied (1.2), p<0.001) for hip and knee surgery] and [(2.1 (1.4) not satisfied versus 2.6 (1.2) satisfied , p<0.001 for back surgery]); and were significantly higher for patients who made informed, patient-centered decisions compared to those who did not (2.7 versus 2.3, p<0.001 for hip and knee surgery and 3.2 [0.9] versus 2.0 [1.3], p<0.001 for back surgery).
- For the Valentine et al 2021b paper, a generalized linear and logistic regression model with the General Estimating Equations approach was used to account for clustering of patients within surgeons in a cross-sectional sample to identify relationships between the scale and health outcomes.
  - The developer attests that Valentine et al 2021b's results showed that higher SDM process scores were associated with larger improvements from pre- to post-surgery in mental (b=0.16, p=0.02) and physical health (b=0.25, p=0.02) outcomes for patients who had total joint replacement of the hip or knee but not patients who had spine surgery (all p's greater than 0.26).

- Validity testing was conducted at the accountable-entity level:
  - The developer cites two studies at the site level, and one study (Fowler et al 2021) summarized performance at the group/practice level. The developers noted they tested whether clinical practices that implemented shared decision making had higher SDM scores than sites practicing usual care. The developers used t-tests to compare mean SDM scores from different settings using a Welch's correlation when needed. They also calculated Cohen's d effect sizes for all comparisons. The developer notes that a 0.2 effect size would indicate a small effect, 0.5 indicates a medium effect, and 0.8 indicates a large effect.
    - The developer notes that for osteoarthritis of the knee and hip, patients in the practices where decision aids were used reported significantly better decision processes (2.9 versus 2.5, P <0.001, d=0.49 and 2.9 versus 2.1, P <0.001, d= 0.84, respectively).
    - The developer notes that the difference in the SDM Process Scores for spine practices that did and did not use decision support (3.0 versus 2.75, P=0.12,

d=0.22) was in the expected direction but was not large enough to reach statistical significance.

- Lastly, the developer notes that with regard to breast cancer practices, the practice that had formal decision support had significantly better scores than cancer practices without any decision support interventions (2.7 versus 2.3, P< 0.05, d=0.47).

- The developers also presented content validity from the Valentine et al 2021a paper, which noted that patients were unable to adequately describe shared decision making and their general desire to rate their clinicians highly, which proved to be problematic as the patients lacked a frame of reference for evaluating decision making. The developer notes that the findings resulted in the SDM process survey's focus on clinical decision and on the report of events or behaviors.
- In the previous submission, the developer compared the aggregate SDM Process Score from patients treated at clinical sites that have committed to shared decision making with reports of national cross-sections of patients from the TRENDS survey who made the same decisions and compared the mean SDM scores for four breast cancer clinical sites where three used usual care and one used decision aids. They also compared the mean SDM scores for hip and knee replacement sites that used usual care versus decision aids. Lastly, they compared SDM scores for a clinical site for patients who discussed treatment benign prostatic hyperplasia (BHP) for before the use of decision aids and after the use of decision aids.
    - The developer states that the results indicate that clinical sites who commit to improved decision making attain average scores from their patients that are higher than the average.

Exclusions

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

Risk Adjustment

- The measure is not risk-adjusted or stratified.

Meaningful Differences

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

practice was low due to already using some decision aids and the effect size representing the incremental improvement.

- Overall, the developer suggests that a meaningful difference in scores corresponds to an effect size of at least 0.4 SD.

Missing Data

- In sample 4, there were no missing data for the SDM Process score.
- In sample 5, 1 percent of responders skipped one or more items on the scale. Of those with missing responses, two responders skipped all items on the scale and five responders skipped one item on the scale.
- In sample 6, the online administration did not allow responders to skip questions, so there were no missing data.
- The developers then compared responses to nonresponders and those with and without missing responses from samples 5 and 6 using t-tests or chi square. The developer notes that when comparing responders and nonresponders, there was no difference between gender, site, or race/ethnicity. However, they did find statistically significant differences by age. Additionally, the developers noted that when comparing those with missing data and without missing data, there were no differences between age, race/ethnicity, gender, clinical topic, or site.
- The developer recommends excluding those with one or more missing responses to the survey, considering missing responses did not have a meaningful impact on scores.

Comparability

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

SMP Concerns

- One SMP member sought clarification regarding the specifications, specifically, how the measure scores are calculated when multiple types of surgeries (hip, knee, back) are involved. For example, is the intention to calculate the measure by condition?
- One SMP member questioned whether the patient/encounter level reliability testing was conducted on patients undergoing PCI, as this would be required if they are included in the denominator.
- There were a number of SMP members who were concerned that the accountable entity level reliability testing did not demonstrate adequate reliability for all surgery types (namely, prostate surgery, PCI, and mastectomy).
- SMP member comments about meaningful differences highlighted the concern that it would be useful to see if an SDM measure can differentiate providers who all practice SDM as opposed to those who do and do not and that the information presented does not fully answer the question.
- There were concerns regarding missing data in that nonresponse bias was not explored fully. Specifically, one SMP member pointed out that in sample 5, mean age for responder was 64.5 and while for non-responder mean age was 59.

Appendix B: Additional Information Submitted by Developers for Consideration

Subgroup 1

NQF #3725 Home Dialysis Retention

Measure Developer/Steward: Kidney Care Quality Alliance

Reliability

- **Issue 1 (Reviewer 3, Reviewer 4): The first issue regarding the reliability of the Home Dialysis Retention Measure is the impact of small facility size on reliability estimates.**

    - **Developer Response 1:** We appreciate and understand this issue. This measure only captures *new* home dialysis patients in a measurement year; and only facilities offering/providing home dialysis in the measurement year are captured in the Measure denominator. As a result, the facility size might be an issue when calculating reliability. To account for it, we hypothesize that a rolling-year measure construct might increase measure reliability. As we only had access to a single year of testing data (2021), we opted to test this hypothesis through randomly generating a new "yearly" data for each facility with the assumption that, in the new year, each facility had the same facility size (number of patients had home dialysis) and the same performance on retention of home dialysis for at least 90 days. We combined the 2021 data with the newly simulated yearly data and performed the analysis. As one reviewer (Reviewer 10) commented "*1 year data provided marginal reliability results, but use of two years of data provides sufficient variation and therefore better reliability results.*" The other reviewer (Reviewer 4) commented "*The developers estimated reliability using a beta-binomial model combined with formulas from the Adams RAND tutorial. They initially estimated reliability for a single-year measurement window. In order to estimate reliability for a two-year measurement window, they simulated an additional year of data and re-estimated reliability using the combined real plus simulated data. The 1-year calculation is based on a widely used methodology. The 2-year calculation makes sense intuitively.*"

- **Issue 2 (Reviewer 4): The second issue regarding reliability is the Reviewer's concern that both the single and two-year reliability estimates may be overestimates. ("The developers estimated reliability using a beta-binomial model combined with formulas from the Adams RAND tutorial. They initially estimated reliability for a single-year measurement window. In order to estimate reliability for a two-year measurement window, they simulated an additional year of data and re-estimated reliability using the combined real plus simulated data. The 1-year calculation is based on a widely used methodology. The 2-year calculation makes sense intuitively. Nonetheless, I suspect that both calculations may be over-estimating the true reliability….")**

- **Developer Response 2:** We agree that this issue could be related to the "small facility-specific denominators." As the reviewer pointed out, "*When denominators are small and p's are mis-estimated, an individual provider's true error variance can either be under- or over-estimated. However, on average, across all providers the tendency is to under-estimate the within-provider error variance. This leads to systematically inflated estimates of reliability.*" We agree with this reviewer's comment. We also accept this reviewer's suggestion and approach to avoid the overestimate problem. We recalculated reliabilities for both single year and double year data using the new reliability formula when sample size is small. See below Table 1 (for single year data), Table 2 (for combined single year and one simulated data), and Table 3 (for combined single year and 2 simulated data). The new results confirmed this reviewer's concern that the reliability may be overestimated for small facilities using the method in our submission:
  - For 2021 single year data, the Q1, median, and Q3 reliabilities (vs in submission) are 0.1218 (vs 0.2740), 0.2444 (vs 0.5473), and 0.3753 (vs 1.000).
  - For combined 2021 and simulated data, the Q1, median, and Q3 reliabilities (vs in submission) are 0.5840 (vs 0.7862), 0.7661 (vs 0.9313), and 0.8588 (vs 1.000).

Table 1. Recalculated reliability using 2021 single year data

Distribution of facility size	N of patients	Reliability if each facility had a sample size N^
*	*	*Alpha=17.6811*
*	*	*Beta=3.9594*
Min	1	0.0442
10th	1	0.0442
Q1 (25th)	3	0.1218
Median	7	0.2444
Q3 (75th)	13	0.3753
90th	21	0.4925
Max	157	0.8789

*Cell intentionally left blank.

Table 2. Recalculated reliability using combined 2021 and simulated data (2-year rolling data)

Distribution of facility size	N of patients	Reliability if each facility had a sample size N^
*	*	*Alpha=3.3794*
*	*	*Beta=0.8953*
Min	2	0.3187
10th	2	0.3187
Q1 (25th)	6	0.5840
Median	14	0.7661

Distribution of facility size	N of patients	Reliability if each facility had a sample size N^
Q3 (75th)	26	0.8588
90th	42	0.9076
Max	314	0.9866

*Cell intentionally left blank.

Table 3. Recalculated reliability using combined 2021 and 2 simulated data (3-year rolling data)

Distribution of facility size	N of patients	Reliability if each facility had a sample size N^
*	*	*Alpha=1.8730*
*	*	*Beta=0.5713*
Min	3	0.5510
10th	3	0.5510
Q1 (25th)	9	0.7864
Median	21	0.8957
Q3 (75th)	39	0.9410
90th	63	0.9627
Max	471	0.9948

*Cell intentionally left blank.

^ Corrected reliability for small sample size: reliability = (squared correlation between p and p-hat) = 1/[1+(alpha+beta)/N].

The above findings revealed by these new analyses indicate that a 3-year rolling average is more appropriate for this measure; we amend the specifications as such. Likewise, consistent with CMS's exclusion rules within the federal ESRD programs, we recommend that facilities that treat <11 patients during the performance period be excluded from the measure.

Other General Comments

Describe any additional information or considerations (that may not be related to reliability or validity) you would like the SMP to be aware of as they reconsider your measure.

- **Issue 3 (Reviewer 8): "The sampling appears to be done at the facility level. However, the application states that "if a company (dialysis organization) owns multiple facilities in a given Hospital Referral Region (HRR) it would report an aggregated score for all facilities located within the HRR owned wholly or in part by the company." The primary sampling unit is therefore unclear. Conflating ownership with facility biases the attribution of results. Also, what does owned "in part" mean? Who are the other owners and what does proportion of ownership imply regarding consistency of practices that are quality based? The specifications are very confusing and no data (e.g. a CONSORT diagram) are provided for exclusions (see p. 13)."**

- **Developer Response 3:** We tested the measure at the facility-level. This level was necessary because the retention measure is dependent on a facility providing home dialysis. If the HRR unit were used for testing, it would capture facilities that have no home dialysis patients and skew the results in an inappropriate manner and compromise reliability. Thus, even if the measure were to be used in a program that aggregates facilities at the HRR level (which the CMMI ESRD Treatment Choices [ETC] Model does [more information about this model is available https://innovation.cms.gov/innovation-models/esrd-treatment-choices-model]), the measure would be evaluating individual facilities that provide home dialysis within the aggregate groups.

If adopted into the CMMI ETC Model, which is the intent of the measure developer, the Centers for Medicare & Medicaid Services (CMS) would aggregate the scores of these individual facilities that provide home dialysis into their CMS-determined HRRs**.**

The reviewer raised concerns about the conflating ownership with facilities potentially biasing the attribution of results. This concern is unwarranted given the organizational structure of how dialysis, particularly home dialysis, is delivered by the overwhelming number of facilities in the country.

We believe it may be helpful to the review to share more about the organizational structure and why CMS chose to use HRRs to aggregate results. Again, we emphasize that this aggregation does not create a bias that would jeopardize the performance of the measure in terms of the reliability testing conducted by CDRG. This example may also address the reviewer's concerns about how ownership is defined.

Under the current ETC Model using HRRs, facilities with common ownership - in whole or part - are aggregated into a single entity for purposes of the rate measures already in the ETC model. Common ownership is generally in whole because facilities are owned by one of the dialysis organizations, such as DaVita, FMC, Dialysis Clinics Inc. (DCI), U.S. Renal Care, Atlantic Dialysis, etc. Even smaller entities like Atlantic Dialysis have common ownership of several facilities. There are some facilities that may be considered to be owned in part. These facilities tend to be part of a joint venture with physician groups.

Because of this common ownership the policies and procedures that govern behavior at the facilities are greatly centralized under the organizations' Chief Medical Officers. As a result, there is much greater continuity among the practice patterns of facilities with common ownership than there might be in other parts of the health care system.

This common ownership also leads to a unique situation in terms of the delivery of home dialysis and is one of the reasons CMS aggregated facilities by HRRs.

Dialysis organizations in a local area will create specific facilities that specialize in the delivery of home dialysis. This structure means that a patient who may start at facility A and selects home dialysis will be transferred to the organization's home dialysis-specific facility. Facility A may appear to have zero home dialysis patients, but that is only because those patients receive their treatment in a different facility.

For example, let's say a patient lives in Indianapolis, Indiana. There may be six dialysis facilities owned by Organization ABC in the city. One of those may specialize in home dialysis. So any patient who selects home dialysis will receive services from that home dialysis-focused facility. As a result, the other five facilities owned by ABC will have zero home dialysis patients. If evaluated independently, these five facilities would be penalized because when one of their patients wants to select home dialysis, the patient is transferred to the home dialysis-focused facility. CMS acknowledged this reality and thus decided to aggregate the facilities.

When it comes to the retention measure, the issue is that the 5 Indianapolis facilities that do not "provide" home dialysis should not be counted in the measure because they will skew the results because of the structural manner in which home dialysis is delivered.

Given that the aggregation group is built off of organizational protocols and linked to the business structure of cohorting home dialysis patients into one (or sometimes two) facilities in an area, testing the retention measure at the facility level is consistent with the delivery of home dialysis and does not create a bias.

- **Issue 4: One reviewer (Reviewer 8) asked about the definition of ownership, ownership in part, and how ownership affects the consistency of practice.**
    - **Developer Response 4:** CMS defines these terms in the Code of the Federal Register relying on the Security Exchange Act. Essentially, it means that the aggregate facilities have a common ownership, which generally in the context of dialysis facilities means a large, medium, or small dialysis organization owning multiple facilities. The ETC final rule sets forth how CMS has aggregated facilities and has concluded that the common ownership will drive aligned consistency in practice.

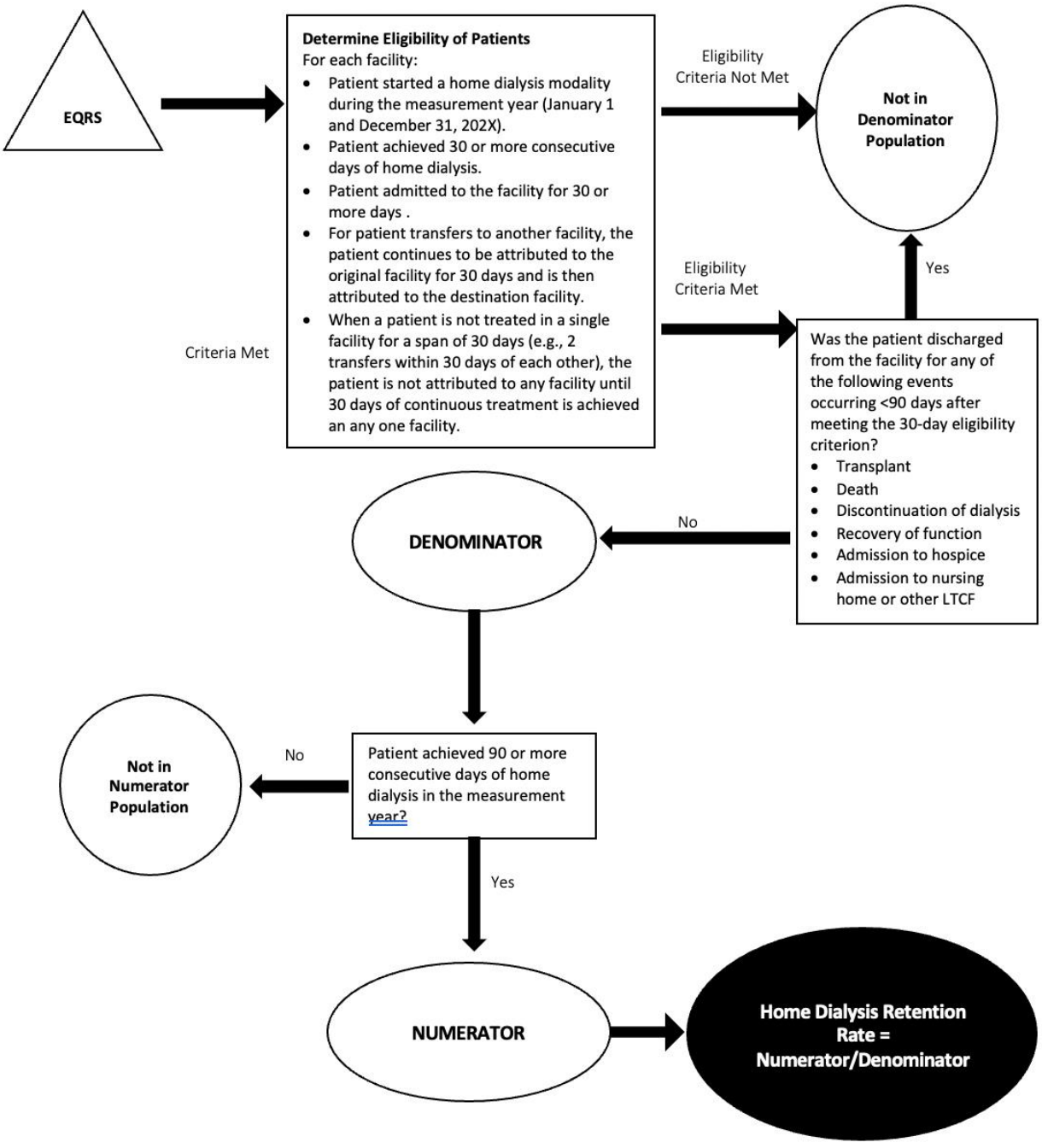
- **Issue 5: Reviewer 8 commented that the specifications were confusing and requested a CONSORT diagram for exclusions.**
    - **Developer Response 5:** We thank the Reviewer for their comment. As noted in the submission documents (see sp. 17 & sp. 18) and specifications table, the Home Dialysis Retention Measure excludes patients who are discharged from the facility <90 days after meeting the denominator eligibility criterion (see below) for any one or more of the following events:

a. Transplant;
b. Death;
c. Discontinuation of dialysis;
d. Recovery of function;
e. Admission to hospice; and/or
f. Admission to nursing home or other LTCF.

The exclusions are intended to avoid potentially disincentivizing or discouraging home dialysis trials by penalizing providers for unanticipated events beyond their realm of control that prevented a patient from achieving the 90-day numerator criterion.

The exclusions were also depicted in the submission documents (sp. 24) in our Calculation Flow Chart Diagram:

HOME DIALYSIS RETENTION CALCULATION FLOW CHART



Denominator Eligibility: As described in the submission documents, to account for the requisite home dialysis training period (up to 4 weeks for home hemodialysis), wherein a certain proportion of patients can be expected to drop out before completion, new home dialysis patients are not eligible for inclusion in the denominator until Day 30 following their first home dialysis treatment, at which time the consecutive time count towards the numerator criterion

commences. The rationale for this "eligibility criterion" is to avoid creating a disincentive for a home dialysis trial by penalizing providers for treatment failures during this training period. (At this point in time, the home dialysis training period cannot be consistently and reliably identified for all patients/payers using administrating or electronic clinical data.)

NQF #3654 Hospice Care Index

Measure Developer/Steward: Abt Associates/Centers for Medicare & Medicaid Services

Reliability

- **Issue 1: Lack of signal-to-noise test for reliability**
    - **Developer Response 1:** For previous endorsement/re-evaluations submissions – most recently for NQF #3645 (Hospice Visits in the Last Days of Life), we followed the methodological approach to reliability testing outlined in "The Reliability of Provider Profiling: A Tutorial" by John Adams of the RAND Corporation (2009) which was featured in the NQF document "What Good Looks Like" for measure submission examples. The actual approach entails using a hierarchical model to obtain an estimate of provider-to-provider variance, and then applying that estimate to the reliability formula (along with estimates of individual provider error). This approach calculates what's known as the "signal-to-noise ratio"; which as Adams (2009) writes, presents "...the proportion of variability in measured performance that can be explained by real differences in performance." I.e., it indicates the extent it can confidentially be ascertained the measure distinguishes performance of one facility. However, this approach is not applicable given the construction of the Hospice Care Index: the reliability formula as one of its inputs requires an estimate of within-provider variation. The index is a hospice-level score, only – i.e., there are not individual-level scores; since the index score relies on a hospice's standing in the national distribution – there is no way to calculate the reliability formula and thus no signal-to-noise ratio possible to be calculated. As the "reliability" concept seeks to ascertain consistency in scoring, we reviewed the stability of scores of hospices over time to measure reliability. During the timeframe of the data we analyzed for testing (2017-2019), the Hospice Care Index was not part of CMS's Hospice Quality Reporting Program, and thereby hospices would be less likely to be actively seeking to improve their performance on its indicators. As hospices were thus acting under typical courses of business, the fact that scores did not vary much from one year to another shows evidence of reliability: hospices presumably acting in the same way received the same score in different years, as we found that index scores were generally stable for a given hospice provider.
- **Issue 2: Need to show data on extent of variation (generally and below/above the threshold) in order to measure reliability**
    - **Developer Response 2:** This is not something we had previously calculated. As with Issue 11 (under Validity), we believe the reviewer is concerned that our approach relies too much on relative differences in hospices (comparing the top 90% vs the bottom 10%). The concern would be that actual differences between the two groups (above and below threshold) might not be meaningful. In developing the

measure, the reality is there are no clinical standards for the domains or indicators represented here. The indicators were drawn from reports from sources such as the Office of Inspector General, MedPAC, and the academic literature, who raised issues, but no evidence exists on what constitutes a meaningful difference in the scores of these indicators. Lacking definite guidelines, our approach was agnostic and focuses on relative comparisons between hospices.

- **Issue 3: Stability analysis did not use any statistical testing – 46% of scores were the same in 2017 and 2019 versus 15% having scores that differed by 2 points or more. Reviewer was not convinced this was sufficient.**
    - **Developer Response 3:** The stability findings that we reported were more detailed than previous stability analyses we had previously seen for measures, which only relied on line graphs (over time). We were not aware of a clear statistical standard for stability and so relied on a descriptive analysis; if there is a particular test or standard we would be happy to explore.
- **Issue 4: Reliability of the 10 individual components of the composite was not assessed**
    - **Developer Response 4:** We looked at this this early on in the development process, and some indicators fared better than others in terms of reliability. For NQF submission we focused on the metrics of the overall index. As helpful we could provide further reliability statistics in the future.
- **Issue 5: Use test-retest reliability testing**
    - **Developer Response 5:** In the matrix, the responder who suggested using test-retest reliability testing points out that a reliability score could be calculated if we use intertemporal variance as a source of within-provider variance; this is actually an interesting idea and we thank the reviewer for it; based on examples given we assumed NQF was more interested in results from cross-sectional, individuals-within-facility variation. The variation over time measures something different, but it seemingly could be something that we could calculate if acceptable to the panel (even with the limitations of our construct). It actually is somewhat related to the stability testing we did do, although it would benchmark a given facility against overall stability – the reviewer notes (and we agree) that stability is imperfect as some hospices could purposefully change due to quality improvement efforts. What is true for the testing timeframe, at least, is that none of the indicators of the index were quality metrics, so quality improvement would not be expected, and changes would be due to true drift. With the assumption no hospice is trying to change their score, we see here the scores are generally stable from year to year.

Validity

- **Issue 1: Consider both top and bottom scores (or all range of scores) instead of scores of 7 and below**
    - **Developer Response 1:** The bottom scores were exceedingly rare: only 13% of hospice scored a 7 or below (2b.06). The very small/rare numbers for the lowest scores confounds that ability to make meaningful statistical comparisons. For this reason, we grouped the lowest scores (which occur most infrequently) together.
- **Issue 2: Correlations with CAHPS Star Rating are of the opposite direction**

- **Developer Response 2:** A higher Hospice Care Index and CAHPS hospice scores both mean higher quality. The Pearson's correlation coefficients that we reported in 2b.03 were positive, indicating the two scores move in the same direction. In that same section, we reported a categorical comparison that perhaps might have been confusing with CAHPS Star Ratings (where again, a higher score is better): relative to a hospice with an HCI score of 10 (the highest), a hospice with a score of 7 or below is almost twice as likely to receive a CAHPS Hospice star rating of one or two – i.e. if a Hospice Care Index score is low, it's more likely that the CAHPS Score will also be low. Our evidence was that the two moved in the same direction.

- **Issue 3: HCI is not risk adjusted, but some indicators in the HCI might warrant risk adjustment (e.g., some patient-level factors would impact things like hospitalizations and visits near death)**
  - **Developer Response 3:** Visits near death specifically is an NQF-endorsed process quality measure (#3645) that is not risk adjusted. That aside, and for the other indicators generally, our general approach was to look at the index holistically. Even if risk adjusted is warranted, we can never be certain if it is done appropriately – or in fact beyond risk adjustment that the indicator is specified to account for all factors: in an indicator like visits near death, there's a possibility fewer happen due to patient family refusals of visits near death: "bad luck" why a hospice would score lower for this indicator. However, this same bad luck should not affect other indicator domains, and the index really seeks to identify hospices registering in outliers for multiple indicator areas simultaneously. A hospice would only be denied a point for being an outlier (typically in the bottom 10% of performance nationwide). Except for those hospices near the threshold, risk adjustment would likely not affect whether a hospice received a point or not.
- **Issue 4: Clarification on the "adjusted odds ratio" since the HCI is not risk adjusted.**
  - **Developer Response 4:** The reviewer is correct that the HCI is not risk adjusted. The "odds ratio" was just the mathematical transformation of a coefficient from a (logistic) regression, to allow more sensible interpretation of the comparison of two categories L specifically, that relative to a hospice with an HCI score of 10 (the highest), a hospice with a score of 7 or below is almost twice as likely to receive a CAHPS Hospice star rating of one or two
- **Issue 5: Low reliability of individual items can pose a potential threat to validity (when scoring is based on ranking of top 90/10% and when denominator sizes can vary)**
  - **Developer Response 5:** This is a valid point, that the hospices the thresholds identify as "outliers" are only arbitrarily identified as such artificially due to small sample sizes. We could investigate further the extent of this. Of course, this problem is common to all measures when dealing with smaller healthcare providers (the situation where this would arise). We hope that this would be mitigated with our index by having multiple indicators; again, if a hospice is identified as an outlier for having a small sample size (and the bad luck of that) there might be other indicators where that isn't true.
- **Issue 6: Application states that there were 4,432 hospices in the sample but only 3,576 hospices were used for the reliability testing. An explanation for the reduction or characteristics of the excluded facilities is necessary.**

- **Developer Response 6:** We apologize for any confusion; the smaller sample was unique for the stability analysis, we only used those hospices large enough to have enough claims to calculate scores in 2017 (only) and 2019 (only), and they also had to appear in both years; i.e., new entrants since 2017 or hospices that terminated service were excluded from our analysis. We acknowledge that this somewhat skews our sample, especially in in favor of larger hospices, compared to the 4,432.

- **Issue 7: Arbitrary dichotomization of HCI into 7 and below vs 10. Scores of 8 and 9 appear to be excluded**
  - **Developer Response 7:** We apologize for any misunderstanding. Scores of 8 and 9 were included in the regression analysis sample; however, the comparison we presented was just for hospices with a score of 10 (the highest) vs. 7 and below (the poorest) to accentuate the range and its correlation with CAHPS hospice. Relative to hospices with a score of 10, hospices with a score of 9 are expected to be less to those hospices with a score of 7 and below; especially because the index seeks to identify hospices failing multiple indicators simultaneously, we focused on the comparison of hospices with perfect scores (10) and those missing multiple indicator points (with scores of 7 and below).
- **Issue 8: Pearson correlations with CAHPS Star Ratings are low/weak**
  - **Developer Response 8:** We acknowledge the correlations with CAHPS are weaker here when compared to other measures we've put forward for endorsement. They do at least move in the correct direction (higher index score hospices have better CAHPS scores). The index is also different in that it's a hospice-level aggregate with only eleven possible values, with most hospices scoring highly, so any correlation comparisons would be more diluted than the person-level comparison with CAHPS scores performed previously (where hospices could have scored on a 0-100% scale).
- **Issue 9: One reviewer expected to see a Cronbach's alpha or other correlation indicators outside of Pearson coefficients for composite analysis.**
  - **Developer Response 9:** That might be an interesting approach to look at for the future (we do assume the evidence would be somewhat consistent between the two approaches)
- **Issue 10: Justification for why all 10 indicators are "process" measures and therefore no risk-adjustment is required.**
  - **Developer Response 10:** This issue overlaps with Issue 3, and we refer the SMP to our response there. The indicators are metrics of hospices service utilization and processes, and while we do not think risk adjustment is needed, the index design of the measure seeks to overcome misspecification in any indicator (should it exist).
- **Issue 11: Constructed measure based on relative performance with no data on how big the differences in performance are**
  - **Developer Response 11:** Our understanding of the issue here questions our approach relies too much on relative differences in hospices (90% vs bottom 10%), and that any difference between the two groups (above and below threshold) might not be meaningful. Although this is possible, the reality is, there are no clinical standards for the domains or indicators represented here. The indicators were drawn from reports from sources such as the Office of Inspector General, who issued reports on the topics covered by the index citing quality concerns. Lacking definite guidelines, our approach was thus more agnostic and sought to focus on relative comparisons.

Other General Comments

Thank you for your review. We appreciate the input and are committed to developing useful measures. We welcome any further thoughts to improve this measure to meet NQF standards.

NQF #3726 Serious Illness Survey for Home-Based Programs

Measure Developer/Steward: RAND Corporation

Reliability

- **Issue 1: ICCs are low (Reviewers 8, 10, and 12).**
    - **Developer Response 1:** Prior research has found that ICCs of 0.01 or greater indicate meaningful variation between health care organizations (Lyratzopoulos et al., 2011). Measures with lower ICCs require larger sample sizes to achieve adequate reliability for unit-level reporting. As shown in Table 2a.5 of the submission, six of the seven proposed Serious Illness Survey for Home-Based Programs measures exhibit acceptable program-level reliability of 0.70 or greater at 100 measure respondents. The remaining measure, Overall Rating, nears the threshold at reliability of 0.67 at 100 respondents. Therefore, we recommend a sample size of 100 completed surveys for making comparisons between home-based serious illness programs. Although there is no national registry of home-based serious illness programs, more than half of the programs that participated in the field test of the survey have 100 or more patients in care at a given time. We anticipate that programs may wish to administer the survey on a rolling basis to achieve sample sizes sufficient to make inter-program comparisons.

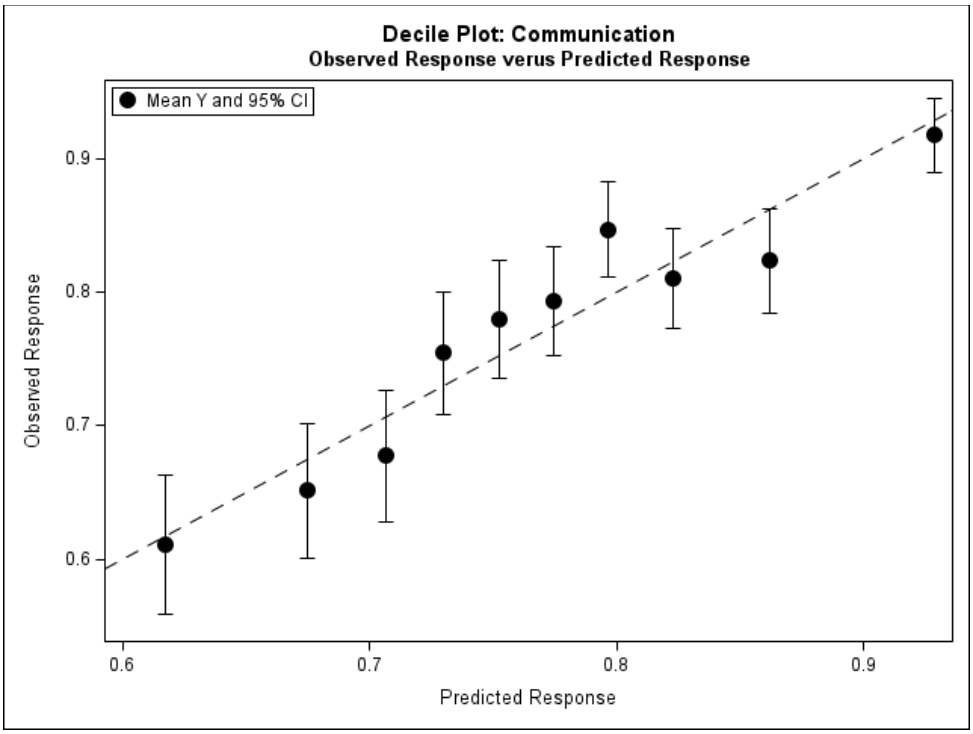
Citation:

Lyratzopoulos G, Elliott MN, Barbiere JM, Staetsky L, Paddison CA, Campbell J, Roland M. August 2011. How can health care organizations be reliably compared? Lessons for a national survey of patient experience. Medical Care. 49(8): 724-733. DOI: https://doi.org/10.1097/mlr.0b013e31821b3482

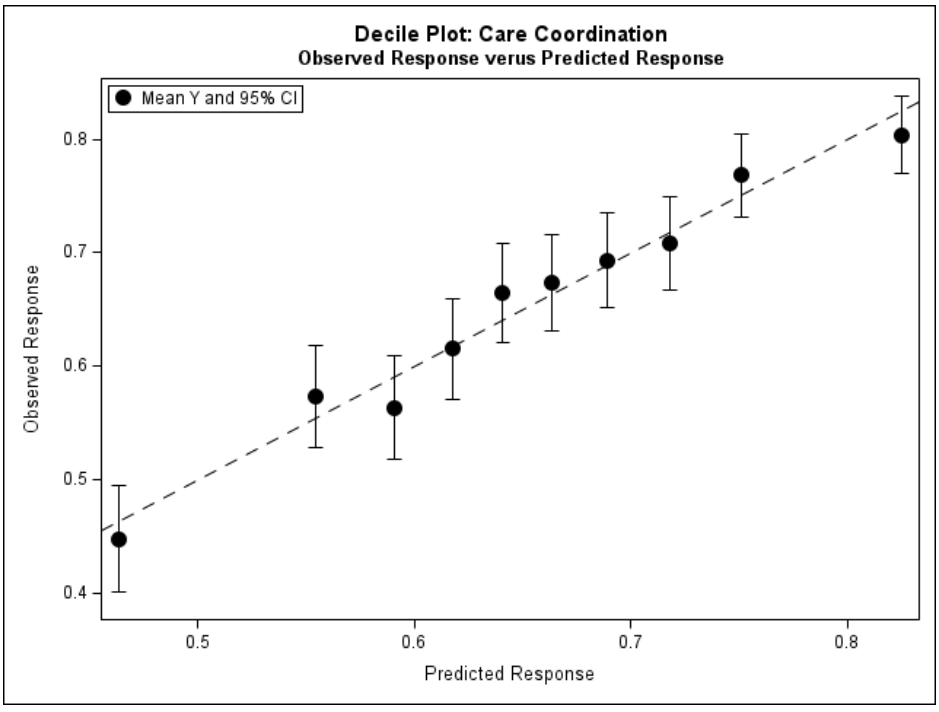
Validity

- **Issue 1: Reviewer 10 requested more information about the overall fit of the risk adjustment model and noted that R-Squared values are modest.**
    - **Developer Response 1:** The following decile plots provide further information regarding the fit of the risk adjustment model. For each proposed measure, the decile plot presents the averaged observed response for each decile of the predicted response. The decile plot includes a diagonal line, which is the line of perfect agreement between the model and the data. In each of the plots, the 10 empirical means of the deciles fall close to the line and also vary randomly above and below the line, indicating that the model is well-specified for each proposed measure.

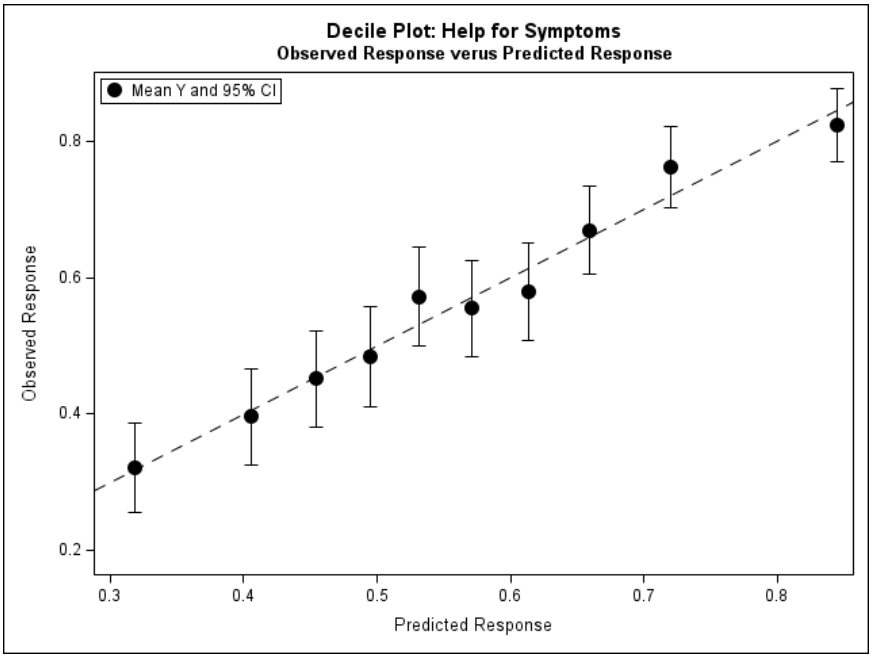
Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Communication Measure



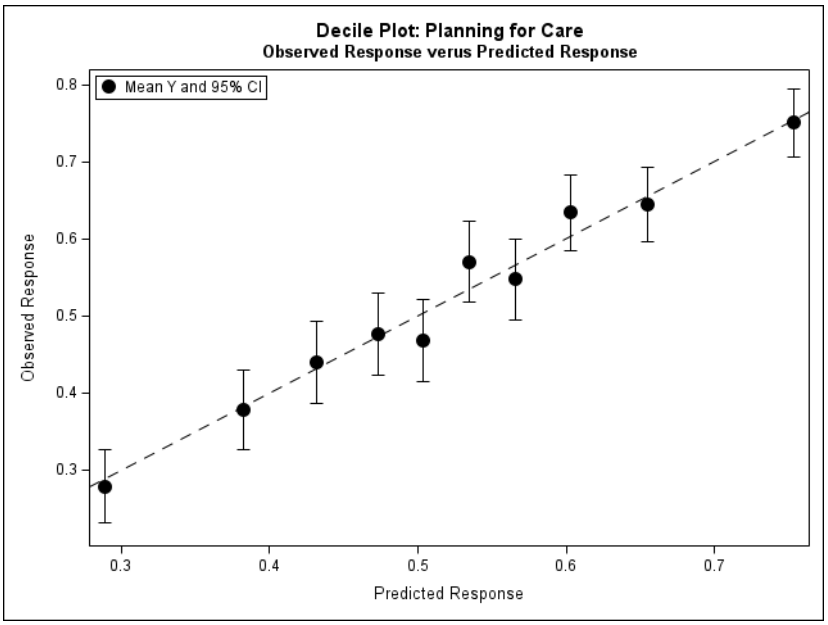
Decile Plot for the Proposed Serious Illness Survey for Home-Basedd Programs Care Coordination Measure



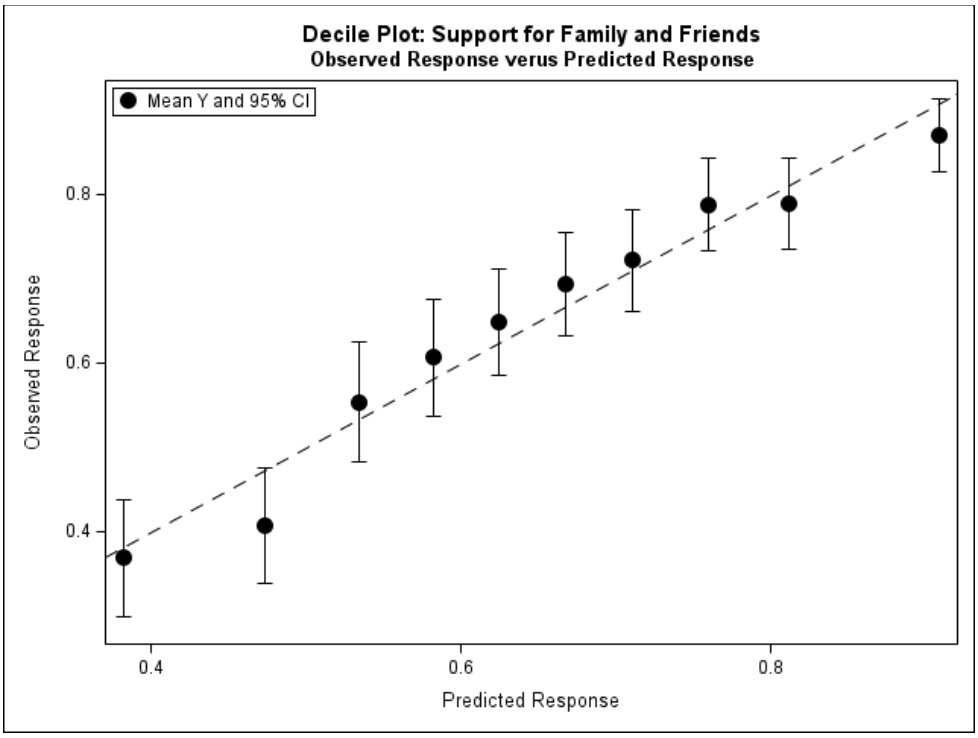
Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Help for Symptoms Measure



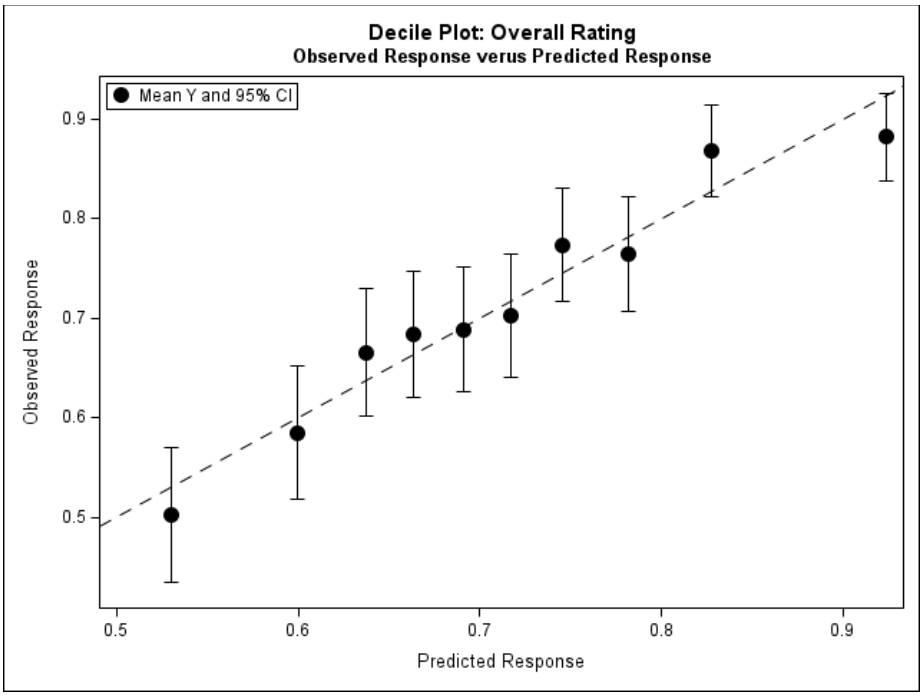
Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Planning for Care Measure



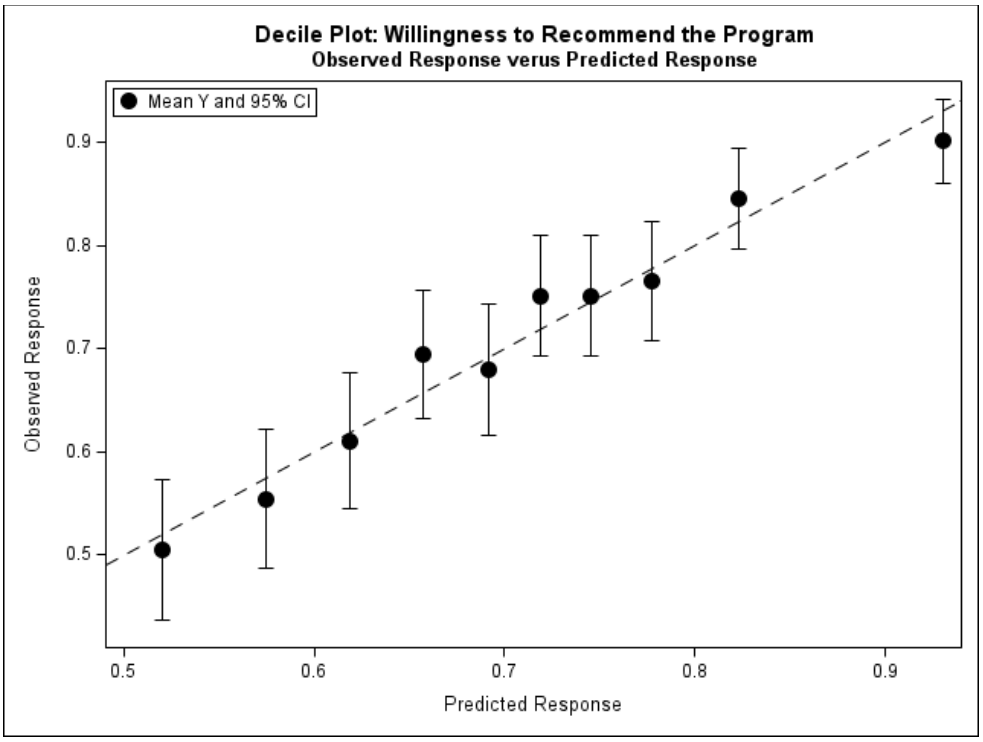
Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Support for Family and Measure



Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Overall Rating Measure



Decile Plot for the Proposed Serious Illness Survey for Home-Based Programs Willingness to Recommend the Program Measure



R-squared values are expected to be more modest for patient experience measures than for clinical outcomes. The r-squared values are similar to those reported for other patient experience measures (see, for example, Zuckerbraun, Owens, et al., 2018), and these r-squares typically decrease as question design improves, because in this context the r-squares reflect patients interpreting questions differently. Well-designed survey items are answered similarly by people with similar experiences. Low within-program disparities can also reduce the r-square. The risk adjustment model accounts for factors that prior research suggests are likely to matter, and accounts for flaws in measurement equivalence in the unadjusted scores. Modest R-square values suggests that these flaws may be relatively small; however, case-mix adjustment is still important and valuable for addressing such flaws when calculating and comparing scores at the level of the reporting unit.

Citation:

Zuckerbraun S, Owens C, Frasier A, Eicheldinger C, Kilpatrick G, Loft JD. Mode and patient-mix adjustment of the inpatient rehabilitation facility experience of care survey. (2018). Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/IRF-Quality-Reporting/Downloads/ModePatientMixPublicFacingIRF.pdf. Last accessed: October 3, 2022.

- **Issue 2: Given the number of programs participating in the field test, it is difficult to make conclusions about whether measure scores can detect meaningful differences between programs (Reviewer 12).**
    - **Developer Response 2:** Prior research suggests a rule of thumb of including data from 25 or more entities to obtain accurate estimates of group-level characteristics (Bryan and Jenkins, 2015). Adhering to this standard, we use data from 28 field test

programs to calculate results regarding meaningful differences in performance (shown in Section 2b.07). In addition, the home-based serious illness programs participating in the field test are broadly representative of the universe of such programs. As shown in Table 2a.1, field test programs provide care across regions of the United States and reflect a range of different sizes and ownership types.

- As described in Section 2b.07, across measures from the Serious Illness Survey for Home-Based Programs, between 18 and 29 percent of programs participating in the field test scored either significantly above or below the field test program average. This indicates that the measures can identify statistically significant differences in programs' performance. Among program scores that were significantly above or below the average, the mean absolute difference between the programs' scores and the average program score for a given measure ranged from 11.0 for Communication to 17.9 for Care Planning, indicating large differences from the program average.

Citation:

Bryan ML and Jenkins SP. Multilevel modelling of country effects: a cautionary tale. (2016). *European Sociological Review*. 32(1): 3-22.

- **Issue 3: Reviewer 3 suggested that Spanish language be considered for inclusion in the case-mix adjustment method.**
  - **Developer Response 3:** The recommended set of case-mix adjustors includes variables that meet the following criteria: they were statistically significantly associated with respondent evaluations ($p < 0.01$) and have an impact ($1 - r$-squared) of at least 1% for one or more outcome measures. These criteria are markers of whether a candidate variable for adjustment has a notable association with patient-reported outcomes. Spanish language did not meet these criteria, so was not recommended for inclusion in the final set of case-mix adjustors (DeYoreo et al., 2022).

Citation:

DeYoreo M, Anhang Price R, Montemayor CK, Tolpadi A, Bradley MA, Schlang D, Teno JM, Cleary PD, and Elliott MN. 2022. Adjusting for patient characteristics to compare quality of care provided by serious illness programs. *J Pall Med.* 25(7).

- **Issue 4: Reviewer 10 requested to see validity testing at the level of the accountable entity.**
  - **Developer Response 4:** The following table displays the program-level correlations between multi-item Serious Illness Survey for Home-Based Programs measures and the global rating measures (Overall Rating and Willingness to Recommend the Program), using data from the 28 field test programs with 10 or more completed surveys. In keeping with the individual-level validity findings reported in Table 2b.1, these results indicate moderate to large associations between the multi-item measures and global measures.

Table. Program-level Correlations between Multi-Item Serious Illness Survey for Home-Based Programs Measures and Global Rating Measures.

Measure	Overall Rating of Program	Willingness to Recommend Program
Communication	0.70	0.70
Care Coordination	0.69	0.72
Help for Symptoms	0.79	0.44
Planning for Care	0.77	0.50
Support for Family and Friends	0.60	0.61

*p<0.001

- **Issue 5: Reviewer 3 requested more information on how adjustments are made for mode of survey administration.**
    - **Developer Response 5:** To adjust for mode of survey administration, we recommend including an indicator or dummy variable for survey mode (e.g., an indicator of mail with telephone follow-up) along with the other recommended case-mix adjustors in regression models for survey measure scores. This way, the estimated program-level scores from the fitted model represent the case-mix and mode-adjusted scores.

Other General Comments

None.

NQF #2651 CAHPS® Hospice Survey, Version 9.0

Measure Developer/Steward: Centers for Medicare & Medicaid Services

Reliability

- **Issue 1: Accountable-Entity Reliability Testing: ICCs are 0.03 or lower, raising concerns about hospice-level reliability (Reviewer 8).**
    - **Developer Response 1:** The estimated ICCs for the proposed CAHPS Hospice Survey measures range from 0.012 to 0.030. Magnitude heuristics for practically important differences in health care experiences suggest that ICCs of 0.01 or greater can indicate meaningful variation between units (Lyratzopoulos et al. 2011).
    - For measures where ICCs are closer to the bottom of this range (i.e., close to 0.01), sufficient sample sizes are needed to achieve adequate reliability for unit-level reporting. As shown in Table 2a.5 of the submission, all nine proposed measures exhibit acceptable hospice-level reliability of 0.70 or greater at the expected average number of completed surveys per hospice observed in national implementation of the survey.

Citation:

Lyratzopoulos G, Elliott MN, Barbiere JM, Staetsky L, Paddison CA, Campbell J, Roland M. August 2011. How can health care organizations be reliably compared? Lessons for a national survey of patient experience. Medical Care. 49(8): 724-733. DOI: https://doi.org/10.1097/mlr.0b013e31821b3482

Validity

- **Issue 1: Risk adjustment model was not re-calculated for the submission (Reviewers 5, 6, 10, 11, 12).**

Developer Response 1: There are two reasons that this submission relies upon the CAHPS Hospice Survey case-mix adjustment model in current use for national implementation of the survey, rather than a model re-estimated using data from the 2021 mode experiment.

First, there is no compelling reason to believe that the best set of case-mix adjustors should be different for the revised measures. The revised CAHPS Hospice Survey measures are similar in content to the existing measures. Of the 21 evaluative survey items that compose the revised measures, just three are new. (One of these, regarding hospice care training, is a summary of several training items on the current survey.) Typically, the same case-mix adjustment model is applied to all evaluative items, with coefficients differing across survey items, as there are advantages of consistency to doing so. CMS re-estimates case-mix coefficients using national implementation data every quarter and sees stability in importance of these adjustors over time. These datasets are the best sources of information for case-mix coefficients that apply to all hospices, given the sample sizes involved.

Given the very large sample sizes in national survey data, there is little downside to including a given adjustor for a new item if it is not necessary; this is the same consideration that underlies the current practice of including an adjustor for all measures even when it is particularly predictive only for a subset of measures. On the flip side, there are no new candidate variables that might be important to include as adjustors for the new items.

Second, as noted above, mode experiment data provide less precise estimates of case-mix coefficients than data from national implementation, as mode experiment data are from a representative but much smaller number of hospices and caregivers than national implementation data, which are collected from caregivers from thousands of hospices each quarter. Upon national implementation of the revised survey, CMS will re-estimate case-mix coefficients for the CAHPS Hospice Survey on a quarterly basis using the latest data.

(Of note, randomized mode experiments are critical for valid estimates of survey mode effects because survey mode is selected at the hospice-level; calculating *mode* adjustments based on national implementation data would generate biased estimates. In contrast, there is nothing uniquely advantageous about using mode experiment data for estimating a case-mix model as case-mix adjustment does not require randomization.)

- **Issue 2: There are large mode effects (Reviewer 4); survey mode adjustment should be required (Reviewers 5 and 12).**

Developer Response 2: As noted in Section 2b.14 of the submission, CMS agrees that mode adjustment is important, and adjusts CAHPS Hospice Survey measure scores for mode of survey administration to ensure that scores are comparable across hospices regardless of the survey mode selected by the hospice.

- **Issue 3: More information is needed regarding non-response (Reviewer 10) and exclusions (Reviewers 8 and 12, with Reviewer 8 calling out exclusions for no-publicity in particular), and the degree to which these contribute to response bias.**

Developer Response 3: The estimated response rate for the revised CAHPS Hospice Survey administered during the 2021 mode experiment was 31.5 percent in Telephone Only mode, 35.1 percent in Mail Only mode, 39.7 percent in Web-Mail mode, and 45.3 percent in Mixed Mode (mail with telephone follow-up). These rates are higher than those observed for other care experience surveys in national use. Caregivers of older decedents were more likely to respond than younger decedents;

caregivers of Black and Hispanic decedents were less likely to respond than caregivers of White decedents; caregivers of male decedents were less likely to respond than caregivers of female decedents; caregivers of decedents whose final setting of care was a nursing home or acute care hospital were less likely to respond than those whose final setting was at home; and caregivers of decedents with shorter final episodes were less likely to respond than those with longer episodes. Caregivers who were the decedent's spouse/partner and caregivers who were the decedent's parent were more likely to respond than caregivers who were the decedent's child.

Although response propensity varies by certain caregiver and decedent characteristics, previous work in other CAHPS settings has demonstrated that nonresponse weighting to account for potential bias is not needed after case-mix adjustment (see, for example, Elliott, Edwards et al. 2005 and Elliott, Zaslavsky et al. 2009). Case-mix adjustment addresses nonresponse bias with greater statistical efficiency than nonresponse weighting.

Very few decedents/caregivers are excluded due to the no publicity exclusion. In Quarter 4, 2021, for example, the average number of decedents/caregivers excluded for this reason by hospices participating in the CAHPS Hospice Survey was 1.5.

Citations:

Elliott MN, Edwards C, Angeles J, Hays RD (2005). "Patterns of unit and item non-response in the CAHPS® Hospital Survey." *Hlth Serv Res* 40(6): 2096-2119.

Elliott MN, Zaslavsky AM, Goldstein E, Lehrman W, Hambarsoomian K, Beckett MK, Giordano L (2009). "Effects of survey mode, patient mix, and nonresponse on CAHPS Hospital Survey scores." *Hlth Serv Res* 44(2): 501-508.

- **Issue 4: Contextualize findings regarding whether measures can identify meaningful differences in performance by providing evidence of validity of measures for distinguishing performance across hospice or patient characteristics (Reviewer 8).**

Developer Response 4: Prior research using CAHPS Hospice Survey data has found that survey measure scores vary significantly across hospice and patient characteristics. For example:

Hospices are more likely to be in the top quartile of CAHPS Hospice Survey scores if they are not-for-profit and not part of a chain or a government hospice, provide care to fewer than 200 patients per year, and serve a rural area (Anhang Price et al., 2020).

Reported care experiences are significantly worse for decedents who received hospice care in a nursing home compared to those who received hospice care at home for all CAHPS Hospice Survey measures (Quigley et al., 2020).

Reported care experiences differ across racial and ethnic groups. For example, caregivers of Black and Hispanic decedents are less likely to report that they received the right amount of emotional and spiritual support than caregivers of White decedents (Anhang Price et al., 2017).

These publications use national implementation data from the current version of the CAHPS Hospice Survey. Given the similarity between the current and revised versions of the survey, CMS anticipates that the revised version will be able to distinguish differences in hospice care experiences across hospice and patient characteristics, as well.

Citations:

Anhang Price R, Parast L, Haas A, Teno JM, Elliott MN. (2017). Black and Hispanic Patients Receive Good Care, but from Poorer Quality Hospices. *Health Affairs.* 36(7): 1283-1290.

Anhang Price R, Tolpadi A, Schlang D, Bradley MA, Parast L, Teno JM, Elliott MN. (2020). Characteristics of Hospices Providing High-Quality Care. *J Palliat Med.*

Quigley DD, Parast L, Haas A, Elliott MN, Teno JM, Anhang Price R. (2020). Differences in Caregiver Reports of the Quality of Hospice Care Across Settings. *J Am Geriatr Soc.* doi: 10.1111/jgs.16361.

Other General Comments

One reviewer (Reviewer 8) asked why the survey excludes hospice patients who are still alive. The CAHPS Hospice Survey is administered solely to bereaved family caregivers to ensure comparability of the timeframe of assessment (i.e., all hospice care through the end of life, which is inherently impossible for patients to assess during active care), and to reduce the likelihood of selection bias, since many hospice patients are too ill or cognitively impaired to respond to a survey, and many receive care for such a short period of time that it would be infeasible to administer a survey during the course of care.

One reviewer (Reviewer 8) asks for clarification of the score type in Table 2b.1, the table that counts the number of hospices that score significantly above or below the average of the mode experiment's participating hospices. In keeping with all score calculations in the submission, this table uses top-box scores adjusted for survey mode and case mix.

Subgroup 2

NQF #3721 Patient-Reported Overall Physical Health Following Chemotherapy Among Adults With Breast Cancer

Measure Developer/Steward: Purchaser Business Group on Health

Reliability

- **Issue 1:** Accountable-Entity Level Reliability Testing – Reviewers commented on the group-level reliability estimate, the approach to reporting the proportion of groups in the sample with sufficient reliability and the minimum accepted sample size per group.
    - **Developer Response 1:** In conducting reliability testing, we analyzed the overall "signal-to-noise" reliability at the average group size (32 patients per group) for the performance measure as well as the minimum sample size required to obtain a nominal reliability of 0.7. We then estimated the group specific reliability that was calculated using each group's sample size. For this measure, the overall reliability was .53 with a 95% confidence interval of (.10, .92). The group specific reliability ranged from .18 to .70 with a mean of .45. Applying a reliability threshold of 0.60, 50% of groups have reliability that is .60 or greater. Applying a reliability threshold of 0.70, 10% of groups have reliability that is .70 or greater.

      Regarding the comment that a minimum accepted sample size is 66, the minimum accepted sample size for PRO-PMs is often determined using empirical methods such as those described here; for example, the recently NQF-reviewed measure #3665 Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood recommends a minimal sample size of 40 in order to achieve an ICC of .7.

- **Issue 2:** Measure Specifications, Survey Timepoints – Allowable Windows – A reviewer commented that the allowable window for survey administration at baseline could be problematic if patients are already experiencing effects of chemotherapy.
    - **Developer Response 2:** The time windows for survey administration were established with direction from the TEP, which included 11 practicing oncology clinicians (see Table 2 for the PROMOnc Technical Expert Panel (TEP) roster in Other General Comments). Over the course of 5 meetings, the TEP carefully considered balancing clinical meaningfulness of the PROMIS scores with the norms of clinic schedules and workflows. Important differences were discussed between parenteral chemotherapy, administered in the practice infusion setting, and oral chemotherapy, taken in the patients' homes. Oncology providers have full visibility into the oral chemotherapy prescription date; however, the actual start date can be influenced by authorizations, pharmacy delays, and patient timeliness and preferences. Oncology providers are often not able to ascertain the actual start date until the patient returns for a check-in visit. In their deliberations regarding this uncertainty, the TEP broadened the PROMIS administration window for oral chemotherapy to promote patient capture. Another consideration is that most side effects and toxicities of common breast cancer oral chemotherapy agents do not interfere with the measures we collected until after the first week of administration with rare exception.

      The implementation guide for PROMOnc explicitly recognized these challenges with oral chemotherapy. Users were instructed to prioritize PROMIS administration prior to administration and only extend beyond if necessary.

- **Issue 3:** Measure Specifications – Baseline Data – A reviewer commented that it was not clear how baseline PROMIS survey scores are used in the measure calculation.
    - **Developer Response 3:** PROMOnc sought to evaluate breast cancer patients' symptoms as they transitioned from treatment to survivorship phase (see Validity, Developer Response 7 for additional description of the measure rationale). As such, the measure numerator is based on the PROMIS survey scores administered about 3 months after completion of chemotherapy.

      Each patient's baseline PROMIS score provides important information for interpretation of their PROMIS score as they enter survivorship, and adjusting for baseline scores are common (for example, Naughton et al., PROMIS-10 scores at six months post-baseline among breast and gynecologic oncology patients participating in a text-based symptom monitoring program with patient navigation., Journal of Clinical Oncology 2020 38:15_suppl, e19173-e19173). In the PROMOnc PRO-PM calculations, the baseline PROMIS scores are used as risk adjustment variables, in accordance with recommendations by the PROMOnc TEP. During this analysis, each patient's follow up PROMIS scores are adjusted based on their baseline PROMIS scores. This adjustment allows for us to control for patient characteristics at baseline that are not under the control of the group but related to the patient's response to the follow-up survey.

Validity

- **Issue 1:** Face validity – Reviewers commented on the adequacy of face validity given some votes were moderate (e.g., rated a 3 on a 5-point scale) and four oncologists declined to participate in voting due to concerns about the testing sample size and/or impact of COVID; there were also requests for the list of face validity panelists, whether patients were consulted, and a comment that the same text/figure appears in the data element validity sections for #3718, #3720 & #3721 measures and therefore accidental copy & paste may have occurred**.**
    - **Developer Response 1:** The roster of PROMOnc Face Validity Panel experts is listed in Table 3 the Other General Comments section. Per NQF recommendations, face validity was conducted by clinicians who were not members of the TEP or otherwise participants in PROMOnc measures testing. Our intent was to conduct face validity testing predominantly with the measured entity, i.e., clinical oncologists but we also recruited a leader from the American Cancer Society who was retired at the time. These experts were identified through outreach to leadership at the Community Oncology Alliance (COA) and the American Society of Clinical Oncology (ASCO).

The PROMOnc measure developers acknowledge the impact of the COVID public health emergency on our testing efforts. The unfortunate overlap of the public health emergency with some of the PROMOnc testing period caused significant oncology practice disruption and resulted in less robust testing data than anticipated. We appreciate and value the feedback of our independent face validity reviewers, including those who chose to defer their voting until additional PROMOnc measure data are available. We did, however, have sufficient testing data to complete the full analysis presented. As in many measure testing projects, PROMOnc will expand and refine testing analyses during implementation for maintenance submission.

Future maintenance testing will include expanded empirical validity testing, to meet requirements for maintenance submission. As described in the Validity Developer Response 2 below, only initial empirical validity testing was completed during this development process.

Patients and caregivers were engaged throughout the PROMOnc testing process. PROMOnc engaged the Patient and Caregiver Oncology Quality Council from the Michigan Oncology Quality Consortium (MOQC) to provide input into the selection of PROMIS scales for assessing patient-reported outcomes. Two representatives from the MOQC Patient and Caregiver Oncology Quality Council also participated on the PROMOnc Steering Committee. See Table 4 in Other General Comments for the Steering Committee roster. And, PROMOnc collaborated with the Seattle Cancer Care Alliance (SCCA) Patient Family Advisory Council (PFAC) on implementation of a patient burden questionnaire during testing.

Further, PROMIS development and research has been based on active patient engagement, including focus groups to inform the survey development and cognitive interviews of survey questions using feedback from patient focus groups about the

outcome domains to make sure that the questions reflect how potential respondents experience the symptoms and outcomes (see, for example, DeWalt et al. 2007).

Regarding the comment that the same text/figure appears in the data element validity sections for #3718, #3720 & #3721 measures and therefore accidental copy & paste may have occurred, each data element in the PROMOnc data dictionary is used for all three measures and thus the table (Table 2b.1: Data Element Validity Among Patients with Data in PROMOnc and Cancer Registry Datasets) listing the purpose of the data element, the data element, the number of patients, agreement index, sensitivity and specificity is the same for all three measures. (The three measures have a common denominator, denominator exclusions, and risk adjustment model so each data element is used for all three measures.)

Regarding the comment that the same text/figure appears in the section with Statistical Results for Validity (e.g., Patient/Encounter Level Validity Testing, face validity), the explanation of the method is the same for each measure but the results content is different for each measure.

- **Issue 2:** Accountable Entity Level Validity Testing – A reviewer commented that accountable entity level validity results were not provided.
    - **Developer Response 2:** During the testing process, the PROMOnc TEP discussed empirical validity testing; however, we were challenged by the paucity of validated, publicly available quality measure data related to these PRO-PMs. TEP members hypothesized only moderate correlation between the PROMOnc measures and available patient experience measures, for instance. The performance data available for comparison across the PROMOnc test sites also varied based on the practice type; e.g., hospital based sites had CAHPS data available while non-hospital based did not; some sites collected standardized oncology ambulatory surveys while others did not; some sites participate in ASCO's Quality Oncology Practice Initiative (QOPI) while others do not.

    Acknowledging these limitations, we did collect data from test sites during the testing time period for H-CAHPS, Outpatient Oncology Press Ganey (note: different items were used across sites), and QOPI (note: different measures were used across sites). Without viewing submitted data, TEP members rated expected correlation strength between the PROMOnc measures and these available data. We then analyzed correlations for any measure for which the TEP hypothesized a moderate association and for which we had data for at least 7 test sites. The results for these 4 resulting measures are presented in Table 1 below. The correlations are in the moderate range, as hypothesized, and in the appropriate direction.

Table 1: Measure Level Empirical Validity

*	Likelihood of your recommending our services to others (Outpatient Oncology Press Ganey)	Degree to which your care was well coordinated among your caregivers (Outpatient Oncology Press Ganey)	Likelihood of recommending hospital (H-CAHPS)	Overall rating of care (H-CAHPS)
Site Count	10	9	7	7
Physical Health Score	PCC* = 0.351	PCC* = 0.342	PCC* = 0.636	PCC* = 0.502

*Pearson's Correlation Coefficient

* Indicates cell left intentionally blank

Further exploration will be conducted during the maintenance phase, and empirical validity testing will be conducted and submitted for maintenance review. During this time, we hope to identify measure(s) with a hypothesized strong correlation for analysis.

- **Issue 3:** Exclusions: A reviewer commented on the exclusion criterion "patient with recurrence/disease progression".
    - **Developer Response 2:** This data element was defined for reporting in the PROMOnc data dictionary and was reported by sites at the time of the follow-up survey.

- **Issue 4:** Risk Adjustment – Reviewers commented on how group practice effect was accounted for in the risk-adjusted score calculation, the sample size in the dataset used for risk model development was small, and noted errors in Table 2b.3.
    - **Developer Response 4:** Thank you for noting the errors in Table 2b.3. The corrected Table 2b.3 is below. We concede that sample size was impacted by COVID and the ongoing analysis for maintenance will be important for this measure, including re-evaluating the variables in the risk adjustment model. However, analyses indicate that the risk adjustment model performs well. Group practice effect was accounted for in the calculation by including fixed effects for groups in the regression model predicting measure scores with the risk adjustor variables. Adjusted group means are then calculated (e.g., using LSMEANS in SAS).

Corrected Table 2b.3:

Table 2b.3. Regression Coefficients in Risk Adjustment Models – Physical Health

Risk Adjustor	Regression Coefficient	Standard Error	p-value
Baseline PROMIS Score	0.38	0.06	0.00
Surgery Level 1	-0.44	1.57	0.78
Surgery Level 2	0.06	1.57	0.97
Surgery Level 3	-11.60	5.21	0.03
Hispanic	-1.96	1.62	0.23

Risk Adjustor	Regression Coefficient	Standard Error	p-value
Non-Hispanic Black	-2.97	1.34	0.03
Non-Hispanic Asian	-4.49	1.62	0.01
Other Race	1.12	1.66	0.50
Former Smoker	-0.75	0.99	0.45
Current Smoker	-2.34	1.54	0.13
Depression	-2.13	2.56	0.41
Diabetic	-1.46	1.18	0.45
Performance Status	0.53	0.41	0.66
Age	0.00	0.43	1.00
BMI	-0.56	1.94	0.19
Aromatase Inhibitor	0.68	0.41	0.48
Days Between Diagnosis and Follow-Up Survey	0.42	0.97	0.32
Days Between Latest Surgery and Follow-Up Survey	-0.28	0.47	0.55
Radiation Within Two Weeks of Follow-Up Survey	1.63	1.14	0.15

- **Issue 5:** Missing data – Reviewers commented that some risk variables had a high rate of missing values and low overall response rate to the survey.
    - **Developer Response 5:** PROMOnc acknowledges that some risk adjustment variables had higher levels of missing data than desired. Based on the clinical expertise and feasibility assessment of our TEP, and knowledge of the literature in oncology practice trends, PROMOnc believes these data are in fact present for a large number of cases for whom they were captured as missing. Throughout the field of oncology, there is increasing attention on ensuring that critical data elements such as those used in PROMOnc are captured in structured fields that can be easily retrieved from an EHR so feasibility of automated data capture is increasing rapidly. When implemented in the context of a reporting program, we anticipate that missing data will be reduced.

  A reviewer suggested that PROMOnc compute the response rate as 323/877, where 323 is the number of completed surveys, and 877 is the number of patients that were eligible for the follow-up survey after removing patients who met the denominator exclusion criteria. If we use this definition, our response rate is 36.8%. However, we think this rate reflects a combination of survey administration rate ([Total Number of Follow-up Surveys Fielded)]/ [Total Number of Patients in the Target Population – Total Number of Patients Meeting the Denominator Exclusion Criteria]) and survey response rate. We computed the survey response rate following the approach commonly used in patient experience surveys, such as CAHPS for MIPS and CAHPS for Hospice, as below:

Response Rate = (Total Number of Completed Surveys) / (Total Number of Follow-up Surveys Fielded – Total Number of Ineligible Surveys)

The Total Number of Completed Surveys is the total number of surveys for which the respondent answers at least 50 percent (9 items in the follow-up survey), which is a threshold commonly used in patient-reported survey measures, of the questions. Total Number of Ineligible Surveys is the total number of surveys for which it is determined that the patient met the denominator exclusion criteria outlined above in Section Sp.17 and including those that have a language barrier or who had mental/physical incapacity.

The reviewer suggested that we report response rate by site. We computed response rate following the reviewer's definition (% Completed Surveys over Number of Patients Eligible for the Follow-up Survey after Removing Patients Meeting the Denominator Exclusion Criteria), as well as using the approach we illustrated above (% Completed Surveys over Total Number of Follow-up Surveys Fielded after Removing Ineligible Surveys). The site response rates are as follows in Table 2:

Table 2: Response Rate by Site Using Two Computational Approaches

Site	% Completed Surveys over Number of Patients Eligible for the Follow-up Survey after Removing Patients Meeting the Denominator Exclusion Criteria	% Completed Surveys over Total Number of Follow-up Surveys Fielded after Removing Ineligible Surveys
1	14.29	46.67
2	29.33	29.33
3	30.81	43.05
4	38.71	44.86
5	42.14	42.14
6	42.31	100.00
7	45.53	48.70
8	60.00	85.71
9	66.67	71.43
10	90.00	100.00

We anticipate that when the measure is implemented outside of the COVID public health emergency and in the context of a reporting program that the 70% threshold is feasible, which was reinforced by the PROMOnc TEP.

We appreciate the reviewer's comment on non-response weighting. Our analyses indicate that response propensity varies by marital status and insurance. We tested these variables as potential risk adjustors and did not include them in the final adjustment model because they presented little association with the measure score (r-square = .007 for marital status and .02 for insurance). We also conducted robustness checks by including these two variables in the risk adjustment model and found inclusion of these two variables has little impact on the performance measure scores

and reliabilities. Previous work in patient experience of care surveys has demonstrated that nonresponse weighting to account for potential bias is not needed after case-mix adjustment (see, for example, Elliott, Edwards et al. 2005 and Elliott, Zaslavsky et al. 2009). When case-mix adjustment suffices to address nonresponse bias, it generally does so with greater statistical efficiency than nonresponse weighting, resulting in estimates of equal reliability and precision with smaller sample sizes, as in our measure testing, than would be required with nonresponse weighting.

References:

- Elliott MN, Edwards C, Angeles J, Hays RD (2005). "Patterns of unit and item non-response in the CAHPS® Hospital Survey." *Hlth Serv Res* 40(6): 2096-2119.
- Elliott MN, Zaslavsky AM, Goldstein E, Lehrman W, Hambarsoomian K, Beckett MK, Giordano L (2009). "Effects of survey mode, patient mix, and nonresponse on CAHPS Hospital Survey scores." *Hlth Serv Res* 44(2): 501-508.

- **Issue 6:** Meaningful Differences – A reviewer commented that empirically observed differences were small and thus the clinical importance of the small difference is unclear.
    - **Developer Response 6:** The literature in the cancer population has suggested to define meaningful difference as between 3- and 6-point difference on a T-score scale that has a mean of 50 and standard deviation of 10 (Jensen et al., 2017; Yost, 2011). Among group scores that were significantly above or below the average, the mean absolute difference between the group's scores and the overall average was 5.19 points, more than half of the standard deviation (5 points). Results indicate that the PRO-PM measure can discriminate between groups' performance.

References:

- Jensen RE, Moinpour CM, Potosky AL, Lobo T, Hahn EA, Hays RD, Cella D, Smith AW, Wu XC, Keegan TH, Paddock LE, Stroup AM, Eton DT. Responsiveness of 8 Patient-Reported Outcomes Measurement Information System (PROMIS) Measures in Large, Community-Based Cancer Study Cohort. Cancer. 2017 Jan 1;123(2):327-335. doi: 10.1002/cncr.30354. Epub 2016 Oct 3. PMID: 27696377
- Yost, Kathleen J. Yost, David T. Eton, Sofia F. Garcia, David Cella (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. Journal of Clinical Epidemiology, Volume 64, Issue 5.

- **Issue 7:** Measure Rationale – A reviewer commented that the rationale for the quality measure was not included nor what a practice can do to manage the outcome.
    - **Developer Response 7:** The measure logic and importance are included in the full NQF Quality Measure Submission Form (Importance to Measure and Report: Evidence (Outcomes) (1a.01-1a.03). Briefly, the rationale notes that: Many patients who undergo chemotherapy with curative intent experience persistent detriments following treatment. Common persistent symptoms include pain, fatigue and detriments to health-related quality of life. Evidence based practices can manage these symptoms during treatment and position patients better for the survivorship phase. This PRO-PM assesses overall physical health following completion of

chemotherapy administered for adult patients with breast cancer. Data from this measure provides insight into the effectiveness of medical oncologists in helping patients to minimize the persistent impact of their treatments.

Evidence-based clinical guidelines, including from the National Comprehensive Cancer Network (NCCN) and American Society of Clinical Oncology (ASCO), provide relevant screening, assessment, and treatment recommendations.

For example, The NCCN Survivorship Guideline (2022), The ACS/ASCO Breast Cancer Survivorship Care Guideline (ACS/ASCO, 2015), The NCCN Cancer -Related Fatigue Guideline (2022), and The NCCN Cancer -Related Fatigue Guideline (2022).

- **Issue 8:** Reason for Validity Score – A reviewer commented that there were no patients on the Technical Expert Panel (TEP).
    - **Developer Response 8:** When we formed the TEP, there were two patient representatives, one who was formerly in an advocacy role at Patients Like Me and one who was an administrator at MOQC, nurse practitioner and a patient. During the measure development period, Patients Like Me was acquired by United Health Group and the other patient excused herself from the TEP when she transitioned to a new job. Moreover, rather than rely on just the personal experience of a small number of patients on the TEP, we engaged the MOQC Patient and Caregiver Oncology Quality Council several times to provide input on key issues such as the outcomes to be measured and the selection of the PROMIS scales for the PROMOnc survey. The Patient and Caregiver Oncology Quality Council is diverse in terms of age, gender, race/ethnicity, cancer type, LGBTQ+, etc. More information about this council can be found here: **https://moqc.org/moqc/poqc/**.

Other General Comments

Table 2: PROMOnc Technical Expert Panel

Committee Member	Title, Organization
Afsaneh Barzi, MD, PhD	Director, Employer Strategy, Associate Clinical Professor, Department of Medical Oncology & Therapeutics Research, City of Hope
Victoria Blinder, MD, MSc / Robert Daly, MD, MBA	Blinder: Assistant Attending Physician, Breast Medicine Service, Department of Medicine, Immigrant Health and Cancer Disparities Service, Department of Psychiatry and Behavioral Sciences), Memorial Sloan Kettering Cancer Center / Daly: Assistant Attending Physician, Department of Medicine, Thoracic Oncology Service, Memorial Sloan Kettering Cancer Center
Stephen B. Edge, MD	VP Healthcare Outcomes and Policy, Roswell Park Cancer Institute
Karen K. Fields, MD	Medical Director, Clinical Pathways & Value-Based Cancer Care, Moffitt Cancer Center
Jennifer Griggs, MD, MPH, FACP, FASCO	Professor, Dept of Health Management & Policy; Dept of Internal Medicine, Hematology & Oncology; Program Director, MOQC

Committee Member	Title, Organization
Emily Mackler, PharmD	Director, Clinical Quality Initiatives, MOQC
Sally Okun	Director, Policy & Ethics; UnitedHealth Group Research & Development; formerly Vice President Advocacy, Policy & Patient Safety, Patients Like Me
Jorge Nieva, MD	Associate Professor of Clinical Medicine, Keck School of Medicine, USC; USC Norris Comprehensive Cancer Center
Bryce Reeve, PhD	Director, Center for Health Measurement, Dept Population Health Sciences, Duke School of Medicine
Dawn Severson, MD	Medical Director, Henry Ford Cancer Institute-Macomb; Cancer Liaison Physician, Henry Ford Health System; Medical Director, Cancer Survivorship Program, HFCI
Angela Stover, PhD	Assistant Professor of Health Policy and Management, The University of North Carolina at Chapel Hill
Ishwaria M. Subbiah, MD, MS	Assistant Professor of Medicine, Department of Palliative Care and Rehabilitation Medicine, MD Anderson
Susan White, PhD, RHIA, CHDA	Administrator of Analytics, Ohio State University-CCC, James Cancer Hospital
Tracy Wong, MBA	Director, Value and Patient Experience, Seattle Cancer Care Alliance
Finly Zachariah, MD / Vincent Chung, MD, FACP	Zachariah: Assistant Clinical Professor, Department of Supportive Care Medicine, City of Hope Chung: Associate Clinical Professor, City of Hope Department of Medical Oncology

Table 3: PROMOnc Face Validity Panel

Committee Member	Title, Organization
Sanjiv Agarwala, MD	Professor, Temple University, Lewis Katz School of Medicine; President & CMO, Cancer Expert Now, Inc.; Community Oncology Alliance
Lakshmi Aggarwal, MD	Fort Wayne Medical Oncology and Hematology; Board Member, Community Oncology Alliance (COA); COA Patient Advocacy Network (CPAN) Medical Co-Chair
Len Lichtenfeld, MD, MACP	Chief Medical Officer, Jasper Health; Former Deputy Chief Medical Officer, American Cancer Society; Cancer Care Board member
Nadine McLeary, MD, MPH	Senior Physician, Dana-Farber Cancer Institute and Brigham and Women's Hospital; Associate Professor of Medicine, Harvard Medical School Dana Farber, Harvard

Committee Member	Title, Organization
Poorni Manohar, MD	Physician, Fred Hutch; Acting Instructor, Internal Medicine, University of Washington School of Medicine; Physician, UW Medicine; Assistant Professor, Clinical Research Division, Fred Hutch; Founder of the Hematology/Oncology Women's Group
Manali Patel, MD, MPH*	Assistant Professor - University Medical Line, Medicine - Oncology Stanford Medicine; Assistant Professor - University Medical Line, Medicine - Oncology
Debra Patt, MD, PhD, MBA*	Physician, Texas Oncology – Austin Central; Past-Chair, ASCO's Clinical Practice Committee; Secretary, Community Oncology Alliance
Blase Polite, MD, MPH	Professor of Medicine, Hematology and Oncology, UChicago Medicine; Chief Physician, University of Chicago Medicine multispecialty care facility; Deputy Section Chief for Clinical Operations and Executive Medical Director for Cancer Accountable Care; Past-Chair of the American Society of Clinical Oncology (ASCO) Health Disparities Committee and a two-time chair of the ASCO Government Relations Committee (GRC)
Derek Raghavan, MD, PhD, FACP*	President of Atrium Health's Levine Cancer Institute. Since joining the Institute
Frederick Schnell, MD, FACP	Physician, Cancer Center of Middle Georgia; Chief Medical Officer, Community Oncology Alliance (COA); Board of Director member, Georgia CORE (Center for Oncology Research & Education); Former CEO, Central Georgia Cancer Care
Juliana Shapira, MD*	Chief Medical Officer, Regional Cancer Care Associates; Associate Professor of Medicine and Cell Biology, SUNY Downstate Medical Center
John Sweetenham, MD, FRCP, FACP	Professor, Department of Internal Medicine at UT Southwestem Medical Center; Associate Director for Clinical Affairs, UTSW's Harold C. Simmons Comprehensive Cancer Center; Board Chair, National Comprehensive Cancer Network; Fellow, American College of Physicians; Fellow, American Society of Clinical Oncology

*Did not vote

Table 4: PROMOnc Steering Committee

Committee Member	Title, Organization
David Lansky, PhD	Senior Advisor, Former President and CEO, Pacific Business Group on Health (Chair)
Catherine Dodd, PhD, RN, FAAN	Director, City and County of San Francisco Health Service System (Retired)
Diane Drago	Member, MOQC Patient and Caregiver Oncology Quality Council (POQC)

Committee Member	Title, Organization
Jennifer Griggs, MD, MPH, FACP, FASCO	Executive Director, Michigan Oncology Quality Consortium (MOQC)
Michael Harrison	Member, MOQC Patient and Caregiver Oncology Quality Council (POQC)
Corinna Andiel, PhD	Associate Director, Quality and Safety, Memorial Sloan Kettering & Chairperson, Quality Committee, Alliance for Dedicated Cancer Centers (ADCC)
Arif Kamal, MD, MBA, MHS, FACP, FAAHPM	Associate Professor of Medicine, Division of Medical Oncology and Duke Palliative Care, Duke University School of Medicine
Jennifer Malin, MD, PhD	Senior Medical Director, Oncology and Genetics, UnitedHealthcare
Mark McClellan, MD, PhD **(sub. Aparna Higgins)**	Director and Robert J. Margolis, MD Professor of Business, Medicine and Policy, Duke Margolis Center for Health Policy (sub. Policy Fellow, Duke-Margolis Center for Health Policy; Founder & CEO Ananya Health Innovations)

NQF #3720 Patient-Reported Fatigue Following Chemotherapy Among Adults With Breast Cancer

Measure Developer/Steward: Purchaser Business Group on Health

Reliability

- **Issue 1:** Accountable-Entity Level Reliability Testing – Reviewers asked for clarification about the group-level reliability estimate and the approach to reporting the proportion of groups in the sample with sufficient reliability.
    - **Developer Response 1:** In conducting reliability testing, we analyzed the overall "signal-to-noise" reliability at the average group size (32 patients per group) for the performance measure as well as the minimum sample size required to obtain a nominal reliability of 0.7. We then estimated the group specific reliability that was calculated using each group's sample size. For this measure, the overall reliability was .77 with a 95% confidence interval of (.48, .93). The group specific reliability ranged from .39 to .88 with a mean of .66. Applying a reliability threshold of 0.60, 50% of groups have reliability that is .60 or greater. Applying a reliability threshold of 0.70, 50% of groups have reliability that is .70 or greater.

- **Issue 2:** Measure Specifications, Survey Timepoints – Allowable Windows – A reviewer commented that the allowable window for survey administration at baseline could be problematic if patients are already experiencing effects of chemotherapy.
    - **Developer Response 2:** The time windows for survey administration were established with direction from the TEP, which included 11 practicing oncology clinicians (see Table 2 for the PROMOnc Technical Expert Panel (TEP) roster in Other General Comments). Over the course of 5 meetings, the TEP carefully considered balancing clinical meaningfulness of the PROMIS scores with the norms of clinic

schedules and workflows. Important differences were discussed between parenteral chemotherapy, administered in the practice infusion setting, and oral chemotherapy, taken in the patients' homes. Oncology providers have full visibility into the oral chemotherapy prescription date; however, the actual start date can be influenced by authorizations, pharmacy delays, and patient timeliness and preferences. Oncology providers are often not able to ascertain the actual start date until the patient returns for a check-in visit. In their deliberations regarding this uncertainty, the TEP broadened the PROMIS administration window for oral chemotherapy to promote patient capture. Another consideration is that most side effects and toxicities of common breast cancer oral chemotherapy agents do not interfere with the measures we collected until after the first week of administration with rare exception.

The implementation guide for PROMOnc explicitly recognized these challenges with oral chemotherapy. Users were instructed to prioritize PROMIS administration prior to administration and only extend beyond if necessary.

- **Issue 3:** Measure Specifications – Baseline Data – A reviewer commented that it was not clear how baseline PROMIS survey scores are used in the measure calculation.
    - **Developer Response 3:** PROMOnc sought to evaluate breast cancer patients' symptoms as they transitioned from treatment to survivorship phase (see Validity, Developer Response 7 for additional description of the measure rationale). As such, the measure numerator is based on the PROMIS survey scores administered about 3 months after completion of chemotherapy.

    Each patient's baseline PROMIS score provides important information for interpretation of their PROMIS score as they enter survivorship, and adjusting for baseline scores are common (for example, Naughton et al., PROMIS-10 scores at six months post-baseline among breast and gynecologic oncology patients participating in a text-based symptom monitoring program with patient navigation., Journal of Clinical Oncology 2020 38:15_suppl, e19173-e19173). In the PROMOnc PRO-PM calculations, the baseline PROMIS scores are used as risk adjustment variables, in accordance with recommendations by the PROMOnc TEP. During this analysis, each patient's follow up PROMIS scores are adjusted based on their baseline PROMIS scores. This adjustment allows for us to control for patient characteristics at baseline that are not under the control of the group but related to the patient's response to the follow-up survey.

Validity

- **Issue 1:** Face validity – Reviewers commented on the adequacy of face validity given some votes were moderate (e.g., rated a 3 on a 5-point scale) and four oncologists declined to participate in voting due to concerns about the testing sample size and/or impact of COVID; there were also requests for the list of face validity panelists, whether patients were consulted, and a comment that the same text/figure appears in the data element validity sections for #3718, #3720 & #3721 measures and therefore accidental copy & paste may have occurred.
    - **Developer Response 1:** The roster of PROMOnc Face Validity Panel experts is listed in Table 3 the Other General Comments section. Per NQF recommendations, face

validity was conducted by clinicians who were not members of the TEP or otherwise participants in PROMOnc measures testing. Our intent was to conduct face validity testing predominantly with the measured entity, i.e., clinical oncologists but we also recruited a leader from the American Cancer Society who was retired at the time. These experts were identified through outreach to leadership at the Community Oncology Alliance (COA) and the American Society of Clinical Oncology (ASCO).

The PROMOnc measure developers acknowledge the impact of the COVID public health emergency on our testing efforts. The unfortunate overlap of the public health emergency with some of the PROMOnc testing period caused significant oncology practice disruption and resulted in less robust testing data than anticipated. We appreciate and value the feedback of our independent face validity reviewers, including those who chose to defer their voting until additional PROMOnc measure data are available. We did, however, have sufficient testing data to complete the full analysis presented. As in many measure testing projects, PROMOnc will expand and refine testing analyses during implementation for maintenance submission.

Future maintenance testing will include expanded empirical validity testing, to meet requirements for maintenance submission. As described in the Validity Developer Response 2 below, only initial empirical validity testing was completed during this development process.

Patients and caregivers were engaged throughout the PROMOnc testing process. PROMOnc engaged the Patient and Caregiver Oncology Quality Council from the Michigan Oncology Quality Consortium (MOQC) to provide input into the selection of PROMIS scales for assessing patient-reported outcomes. Two representatives from the MOQC Patient and Caregiver Oncology Quality Council also participated on the PROMOnc Steering Committee. See Table 4 in Other General Comments for the Steering Committee roster. And, PROMOnc collaborated with the Seattle Cancer Care Alliance (SCCA) Patient Family Advisory Council (PFAC) on implementation of a patient burden questionnaire during testing.

Further, PROMIS development and research has been based on active patient engagement, including focus groups to inform the survey development and cognitive interviews of survey questions using feedback from patient focus groups about the outcome domains to make sure that the questions reflect how potential respondents experience the symptoms and outcomes (see, for example, DeWalt et al. 2007).

Regarding the comment that the same text/figure appears in the data element validity sections for #3718, #3720 & #3721 measures and therefore accidental copy & paste may have occurred, each data element in the PROMOnc data dictionary is used for all three measures and thus the table (Table 2b.1: Data Element Validity Among Patients with Data in PROMOnc and Cancer Registry Datasets) listing the purpose of the data element, the data element, the number of patients, agreement index, sensitivity and

specificity is the same for all three measures. (The three measures have a common denominator, denominator exclusions, and risk adjustment model so each data element is used for all three measures.)

Regarding the comment that the same text/figure appears in the section with Statistical Results for Validity (e.g., Patient/Encounter Level Validity Testing, face validity), the explanation of the method is the same for each measure but the results content is different for each measure.

- **Issue 2:** Accountable Entity Level Validity Testing – A reviewer commented that accountable entity level validity results were not provided.
    - **Developer Response 2:** During the testing process, the PROMOnc TEP discussed empirical validity testing; however, we were challenged by the paucity of validated, publicly available quality measure data related to these PRO-PMs. TEP members hypothesized only moderate correlation between the PROMOnc measures and available patient experience measures, for instance. The performance data available for comparison across the PROMOnc test sites also varied based on the practice type; e.g., hospital based sites had CAHPS data available while non-hospital based did not; some sites collected standardized oncology ambulatory surveys while others did not; some sites participate in ASCO's Quality Oncology Practice Initiative (QOPI) while others do not.

Acknowledging these limitations, we did collect data from test sites during the testing time period for H-CAHPS, Outpatient Oncology Press Ganey (note: different items were used across sites), and QOPI (note: different measures were used across sites). Without viewing submitted data, TEP members rated expected correlation strength between the PROMOnc measures and these available data. We then analyzed correlations for any measure for which the TEP hypothesized a moderate association and for which we had data for at least 7 test sites. The results for these 4 resulting measures are presented in Table 1 below. The correlations are in the moderate range, as hypothesized, and in the appropriate direction.

Table 1: Measure Level Empirical Validity

*	Likelihood of your recommending our services to others (Outpatient Oncology Press Ganey)	Degree to which your care was well coordinated among your caregivers (Outpatient Oncology Press Ganey)	Likelihood of recommending hospital (H-CAHPS)	Overall rating of care (H-CAHPS)
Site Count	10	9	7	7
Fatigue Score	PCC* = -0.430	PCC* = -0.441	PCC* = -0.509	PCC* = -0.330

* Indicates cell left intentionally blank

*Pearson's Correlation Coefficient

Further exploration will be conducted during the maintenance phase, and empirical validity testing will be conducted and submitted for maintenance review. During this time, we hope to identify measure(s) with a hypothesized strong correlation for analysis.

- **Issue 3:** Exclusions: A reviewer commented on the exclusion criterion "patient with recurrence/disease progression".
    - **Developer Response 2:** This data element was defined for reporting in the PROMOnc data dictionary and was reported by sites at the time of the follow-up survey.

- **Issue 4:** Risk Adjustment – Reviewers commented on how group practice effect was accounted for in the risk-adjusted score calculation, the sample size in the dataset used for risk model development was small, and noted errors in Table 2b.3.
    - **Developer Response 4:** Thank you for noting the errors in Table 2b.3. The corrected Table 2b.3 is below. We concede that sample size was impacted by COVID and the ongoing analysis for maintenance will be important for this measure, including re-evaluating the variables in the risk adjustment model. However, analyses indicate that the risk adjustment model performs well. Group practice effect was accounted for in the calculation by including fixed effects for groups in the regression model predicting measure scores with the risk adjustor variables. Adjusted group means are then calculated (e.g., using LSMEANS in SAS).

Corrected Table 2b.3:

Table 2b.3. Regression Coefficients in Risk Adjustment Models – Fatigue

Risk Adjustor	Regression Coefficient	Standard Error	p-value
Baseline PROMIS Score	0.43	0.06	0.00
Surgery Level 1	0.43	1.96	0.83
Surgery Level 2	-1.22	1.95	0.53
Surgery Level 3	8.77	6.56	0.18
Hispanic	1.40	2.06	0.50
Non-Hispanic Black	0.43	1.70	0.80
Non-Hispanic Asian	3.62	2.05	0.08
Other Race	1.43	2.10	0.50
Former Smoker	0.79	1.23	0.52
Current Smoker	0.57	1.91	0.77
Depression	0.89	3.19	0.78
Diabetic	-1.51	1.48	0.53
Performance Status	1.23	0.51	0.41
Age	-1.03	0.52	0.05
BMI	1.16	2.40	0.03
Aromatase Inhibitor	-2.92	0.51	0.02

Risk Adjustor	Regression Coefficient	Standard Error	p-value
Days Between Diagnosis and Follow-Up Survey	-0.51	1.20	0.32
Days Between Latest Surgery and Follow-Up Survey	1.35	0.59	0.02
Radiation Within Two Weeks of Follow-Up Survey	0.26	1.43	0.85

- **Issue 5:** Missing data – Reviewers commented that some risk variables had a high rate of missing values and low overall response rate to the survey.
    - **Developer Response 5:** PROMOnc acknowledges that some risk adjustment variables had higher levels of missing data than desired. Based on the clinical expertise and feasibility assessment of our TEP, and knowledge of the literature in oncology practice trends, PROMOnc believes these data are in fact present for a large number of cases for whom they were captured as missing. Throughout the field of oncology, there is increasing attention on ensuring that critical data elements such as those used in PROMOnc are captured in structured fields that can be easily retrieved from an EHR so feasibility of automated data capture is increasing rapidly. When implemented in the context of a reporting program, we anticipate that missing data will be reduced.

A reviewer suggested that PROMOnc compute the response rate as 323/877, where 323 is the number of completed surveys, and 877 is the number of patients that were eligible for the follow-up survey after removing patients who met the denominator exclusion criteria. If we use this definition, our response rate is 36.8%. However, we think this rate reflects a combination of survey administration rate ([Total Number of Follow-up Surveys Fielded)]/ [Total Number of Patients in the Target Population – Total Number of Patients Meeting the Denominator Exclusion Criteria]) and survey response rate. We computed the survey response rate following the approach commonly used in patient experience surveys, such as CAHPS for MIPS and CAHPS for Hospice, as below:

Response Rate = (Total Number of Completed Surveys) / (Total Number of Follow-up Surveys Fielded – Total Number of Ineligible Surveys)

The Total Number of Completed Surveys is the total number of surveys for which the respondent answers at least 50 percent (9 items in the follow-up survey), which is a threshold commonly used in patient-reported survey measures, of the questions. Total Number of Ineligible Surveys is the total number of surveys for which it is determined that the patient met the denominator exclusion criteria outlined above in Section Sp.17 and including those that have a language barrier or who had mental/physical incapacity.

The reviewer suggested that we report response rate by site. We computed response rate following the reviewer's definition (% Completed Surveys over Number of Patients Eligible for the Follow-up Survey after Removing Patients Meeting the Denominator

Exclusion Criteria), as well as using the approach we illustrated above (% Completed Surveys over Total Number of Follow-up Surveys Fielded after Removing Ineligible Surveys). The site response rates are as follows in Table 2:

Table 2: Response Rate by Site Using Two Computational Approaches

Site	% Completed Surveys over Number of Patients Eligible for the Follow-up Survey after Removing Patients Meeting the Denominator Exclusion Criteria	% Completed Surveys over Total Number of Follow-up Surveys Fielded after Removing Ineligible Surveys
1	14.29	46.67
2	29.33	29.33
3	30.81	43.05
4	38.71	44.86
5	42.14	42.14
6	42.31	100.00
7	45.53	48.70
8	60.00	85.71
9	66.67	71.43
10	90.00	100.00

We anticipate that when the measure is implemented outside of the COVID public health emergency and in the context of a reporting program that the 70% threshold is feasible, which was reinforced by the PROMOnc TEP.

We appreciate the reviewer's comment on non-response weighting. Our analyses indicate that response propensity varies by marital status and insurance. We tested these variables as potential risk adjustors and did not include them in the final adjustment model because they presented little association with the measure score (r-square = .007 for marital status and .02 for insurance). We also conducted robustness checks by including these two variables in the risk adjustment model and found inclusion of these two variables has little impact on the performance measure scores and reliabilities. Previous work in patient experience of care surveys has demonstrated that nonresponse weighting to account for potential bias is not needed after case-mix adjustment (see, for example, Elliott, Edwards et al. 2005 and Elliott, Zaslavsky et al. 2009). When case-mix adjustment suffices to address nonresponse bias, it generally does so with greater statistical efficiency than nonresponse weighting, resulting in estimates of equal reliability and precision with smaller sample sizes, as in our measure testing, than would be required with nonresponse weighting.

References:

- Elliott MN, Edwards C, Angeles J, Hays RD (2005). "Patterns of unit and item non-response in the CAHPS® Hospital Survey." *Hlth Serv Res* 40(6): 2096-2119.
- Elliott MN, Zaslavsky AM, Goldstein E, Lehrman W, Hambarsoomian K, Beckett MK, Giordano L (2009). "Effects of survey mode, patient mix, and nonresponse

on CAHPS Hospital Survey scores." *Hlth Serv Res* 44(2): 501-508.

- **Issue 6:** Meaningful Differences – A reviewer commented that empirically observed differences were small and thus the clinical importance of the small difference is unclear.
    - **Developer Response 6:** The literature in the cancer population has suggested to define meaningful difference as between 3- and 5-point difference on a T-score scale that has a mean of 50 and standard deviation of 10 (Jensen et al., 2017; Yost, 2011). Among group scores that were significantly above or below the average, the mean absolute difference between the group's scores and the overall average was 4.9 points, very close to half of the standard deviation (5 points). These results indicate that the PRO-PM measure can discriminate between groups' performance.

References:

- Jensen RE, Moinpour CM, Potosky AL, Lobo T, Hahn EA, Hays RD, Cella D, Smith AW, Wu XC, Keegan TH, Paddock LE, Stroup AM, Eton DT. Responsiveness of 8 Patient-Reported Outcomes Measurement Information System (PROMIS) Measures in Large, Community-Based Cancer Study Cohort. Cancer. 2017 Jan 1;123(2):327-335. doi: 10.1002/cncr.30354. Epub 2016 Oct 3. PMID: 27696377
- Yost, Kathleen J. Yost, David T. Eton, Sofia F. Garcia, David Cella (2011). Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. Journal of Clinical Epidemiology, Volume 64, Issue 5.

- **Issue 7:** Measure Rationale – A reviewer commented that the rationale for the quality measure was not included nor what a practice can do to manage the outcome.
    - **Developer Response 7:** The measure logic and importance are included in the full NQF Quality Measure Submission Form (Importance to Measure and Report: Evidence (Outcomes) (1a.01-1a.03). Briefly, the rationale notes that: Many patients who undergo chemotherapy with curative intent experience persistent detriments following treatment. Common persistent symptoms include pain, fatigue and detriments to health-related quality of life. Evidence based practices can manage these symptoms during treatment and position patients better for the survivorship phase. This PRO-PM assesses fatigue following completion of chemotherapy administered for adult patients with breast cancer. Data from this measure provides insight into the effectiveness of medical oncologists in helping patients to minimize the persistent impact of their treatments.

Evidence-based clinical guidelines, including from the National Comprehensive Cancer Network (NCCN) and American Society of Clinical Oncology (ASCO), provide relevant screening, assessment, and treatment recommendations.

For example, The NCCN Cancer-Related Fatigue Guideline (2022), The NCCN Survivorship Guideline (2022, page SFAT-1) and The ACS/ASCO Breast Cancer Survivorship Care Guideline (ACS/ASCO, 2015).

Other General Comments

Table 2: PROMOnc Technical Expert Panel

Committee Member	Title, Organization
Afsaneh Barzi, MD, PhD	Director, Employer Strategy, Associate Clinical Professor, Department of Medical Oncology & Therapeutics Research, City of Hope
Victoria Blinder, MD, MSc / Robert Daly, MD, MBA	Blinder: Assistant Attending Physician, Breast Medicine Service, Department of Medicine, Immigrant Health and Cancer Disparities Service, Department of Psychiatry and Behavioral Sciences), Memorial Sloan Kettering Cancer Center / Daly: Assistant Attending Physician, Department of Medicine, Thoracic Oncology Service, Memorial Sloan Kettering Cancer Center
Stephen B. Edge, MD	VP Healthcare Outcomes and Policy, Roswell Park Cancer Institute
Karen K. Fields, MD	Medical Director, Clinical Pathways & Value-Based Cancer Care, Moffitt Cancer Center
Jennifer Griggs, MD, MPH, FACP, FASCO	Professor, Dept of Health Management & Policy; Dept of Internal Medicine, Hematology & Oncology; Program Director, MOQC
Emily Mackler, PharmD	Director, Clinical Quality Initiatives, MOQC
Sally Okun	Director, Policy & Ethics; UnitedHealth Group Research & Development; formerly Vice President Advocacy, Policy & Patient Safety, Patients Like Me
Jorge Nieva, MD	Associate Professor of Clinical Medicine, Keck School of Medicine, USC; USC Norris Comprehensive Cancer Center
Bryce Reeve, PhD	Director, Center for Health Measurement, Dept Population Health Sciences, Duke School of Medicine
Dawn Severson, MD	Medical Director, Henry Ford Cancer Institute-Macomb; Cancer Liaison Physician, Henry Ford Health System; Medical Director, Cancer Survivorship Program, HFCI
Angela Stover, PhD	Assistant Professor of Health Policy and Management, The University of North Carolina at Chapel Hill
Ishwaria M. Subbiah, MD, MS	Assistant Professor of Medicine, Department of Palliative Care and Rehabilitation Medicine, MD Anderson
Susan White, PhD, RHIA, CHDA	Administrator of Analytics, Ohio State University-CCC, James Cancer Hospital
Tracy Wong, MBA	Director, Value and Patient Experience, Seattle Cancer Care Alliance
Finly Zachariah, MD / Vincent Chung, MD, FACP	Zachariah: Assistant Clinical Professor, Department of Supportive Care Medicine, City of Hope Chung: Associate Clinical Professor, City of Hope Department of Medical Oncology

Table 3: PROMOnc Face Validity Panel

Committee Member	Title, Organization
Sanjiv Agarwala, MD	Professor, Temple University, Lewis Katz School of Medicine; President & CMO, Cancer Expert Now, Inc.; Community Oncology Alliance
Lakshmi Aggarwal, MD	Fort Wayne Medical Oncology and Hematology; Board Member, Community Oncology Alliance (COA); COA Patient Advocacy Network (CPAN) Medical Co-Chair
Len Lichtenfeld, MD, MACP	Chief Medical Officer, Jasper Health; Former Deputy Chief Medical Officer, American Cancer Society; Cancer Care Board member
Nadine McLeary, MD, MPH	Senior Physician, Dana-Farber Cancer Institute and Brigham and Women's Hospital; Associate Professor of Medicine, Harvard Medical School Dana Farber, Harvard
Poorni Manohar, MD	Physician, Fred Hutch; Acting Instructor, Internal Medicine, University of Washington School of Medicine; Physician, UW Medicine; Assistant Professor, Clinical Research Division, Fred Hutch; Founder of the Hematology/Oncology Women's Group
Manali Patel, MD, MPH*	Assistant Professor - University Medical Line, Medicine - Oncology Stanford Medicine; Assistant Professor - University Medical Line, Medicine - Oncology
Debra Patt, MD, PhD, MBA*	Physician, Texas Oncology – Austin Central; Past-Chair, ASCO's Clinical Practice Committee; Secretary, Community Oncology Alliance
Blase Polite, MD, MPH	Professor of Medicine, Hematology and Oncology, UChicago Medicine; Chief Physician, University of Chicago Medicine multispecialty care facility; Deputy Section Chief for Clinical Operations and Executive Medical Director for Cancer Accountable Care; Past-Chair of the American Society of Clinical Oncology (ASCO) Health Disparities Committee and a two-time chair of the ASCO Government Relations Committee (GRC)
Derek Raghavan, MD, PhD, FACP*	President of Atrium Health's Levine Cancer Institute. Since joining the Institute
Frederick Schnell, MD, FACP	Physician, Cancer Center of Middle Georgia; Chief Medical Officer, Community Oncology Alliance (COA); Board of Director member, Georgia CORE (Center for Oncology Research & Education); Former CEO, Central Georgia Cancer Care
Juliana Shapira, MD*	Chief Medical Officer, Regional Cancer Care Associates; Associate Professor of Medicine and Cell Biology, SUNY Downstate Medical Center
John Sweetenham, MD, FRCP, FACP	Professor, Department of Internal Medicine at UT Southwestern Medical Center; Associate Director for Clinical Affairs, UTSW's Harold C. Simmons Comprehensive Cancer Center; Board Chair, National Comprehensive Cancer Network; Fellow, American College of Physicians; Fellow, American Society of Clinical Oncology

*Did not vote

Table 4: PROMOnc Steering Committee

Committee Member	Title, Organization
David Lansky, PhD	Senior Advisor, Former President and CEO, Pacific Business Group on Health (Chair)
Catherine Dodd, PhD, RN, FAAN	Director, City and County of San Francisco Health Service System (Retired)
Diane Drago	Member, MOQC Patient and Caregiver Oncology Quality Council (POQC)
Jennifer Griggs, MD, MPH, FACP, FASCO	Executive Director, Michigan Oncology Quality Consortium (MOQC)
Michael Harrison	Member, MOQC Patient and Caregiver Oncology Quality Council (POQC)
Corinna Andiel, PhD	Associate Director, Quality and Safety, Memorial Sloan Kettering & Chairperson, Quality Committee, Alliance for Dedicated Cancer Centers (ADCC)
Arif Kamal, MD, MBA, MHS, FACP, FAAHPM	Associate Professor of Medicine, Division of Medical Oncology and Duke Palliative Care, Duke University School of Medicine
Jennifer Malin, MD, PhD	Senior Medical Director, Oncology and Genetics, UnitedHealthcare
Mark McClellan, MD, PhD **(sub. Aparna Higgins)**	Director and Robert J. Margolis, MD Professor of Business, Medicine and Policy, Duke Margolis Center for Health Policy (sub. Policy Fellow, Duke-Margolis Center for Health Policy; Founder & CEO Ananya Health Innovations)

Appendix C: Measures Withdrawn After SMP Review

Subgroup 1

NQF #2881 Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI)

Reason for withdrawal: Developer will re-evaluate the reliability testing.

Maintenance Measure

Brief Description of Measure: Measure score: The measure is a risk-standardized score at the hospital level for days spent in acute care for patients with an AMI.

Measure focus and time frame: This measure estimates days spent in acute care (i.e., time spent in ED, unplanned readmission and observation stays) within 30 days of discharge from an inpatient hospitalization for acute myocardial infarction (AMI). This measure is intended to capture the quality of care transitions provided to discharged patients hospitalized with AMI by collectively measuring a set of adverse acute care outcomes that can occur post-discharge: 1) emergency department (ED) visits, 2) observation stays, and 3) unplanned readmissions at any time during the 30 days post-discharge. Readmissions are classified as planned and unplanned by applying the planned readmission algorithm

(PRA). Days spent in each care setting are aggregated for the 30 days post-discharge with a minimum of half-day increments (i.e., an ED visit lasting 2 hours would be counted as 0.5 days).

Target population: CMS annually reports the measure for patients who at least 65 years old and enrolled in fee-for-service (FFS) Medicare, and were hospitalized in non-federal hospitals or in Veterans Health Administration (VA) facilities.

Numerator Statement: The outcome of the measure is a count of the number of days a patient spends in acute care within 30 days of discharge from an eligible index AMI hospitalization. We define days in acute care as days spent in an ED, admitted to an observation unit, or admitted as an unplanned readmission for any cause to a short-term acute care hospital, within 30 days from the date of discharge from the index AMI hospitalization.

Denominator Statement: The target population for this measure is Medicare FFS beneficiaries aged 65 years and older hospitalized for AMI at non-federal and VA acute care hospitals. The cohort includes admissions for patients discharged from the hospital with a principal diagnosis of AMI and with continuous 12 months Medicare enrollment prior to admission. The measure is publicly reported by CMS for those patients 65 years and older who are Medicare FFS or VA beneficiaries admitted to non-federal or VA hospitals, respectively.

Denominator Exclusions: The measure excludes index hospitalizations that meet any of the following exclusion criteria:

1. Without at least 30 days of post-discharge enrollment in Medicare FFS
2. Discharged against medical advice
3. Same-day discharges
4. AMI admissions within 30 days of discharge from a prior AMI index admission

Measure Type: Outcome

Data Source: Claims, Medicare Enrollment Data (including Master Beneficiary Summary File), VHA Administrative Data

Level of Analysis: Facility

Risk-Adjusted: Statistical risk model with 31 factors

Sampling Allowed: None

Reliability

Preliminary ratings for reliability: The SMP Did Not Pass on Reliability with a score of: H-0; M-4; L-6; I-1

Specifications:

- Measure specifications are clear and precise.
- Measure specifications have been updated since the SMP's review of this measure in the spring 2021 cycle.
- Relevant evaluation background: This AMI excess days in acute care (EDAC) measure (NQF #2881) was initially submitted to NQF for re-endorsement in the spring 2021 cycle, where it was reviewed by the SMP along with two similar measures that focus on two other conditions (NQF #2880, which focuses on EDAC for patients with heart failure, and NQF #2882, which focuses on pneumonia).

- CMS, the measure steward, withdrew the AMI EDAC measure following a review conducted by the SMP. The measure did not pass on the criterion of reliability and the Standing Committee did not pull the measure for a revote.
- To address reliability concerns, CMS raised the minimum case volume from 25 to 50, and then followed the required regulatory steps for measures with substantive changes, which include submitting the measure to the Measure Applications Partnership (MAP); in June 2022, the MAP Hospital Workgroup voted to support the measure for rulemaking.
- The developer is re-submitting the AMI EDAC measure for NQF re-endorsement; the only change in the measure specification since spring 2021 is the increase in minimum case volume from at least 25 to at least 50.

Reliability Testing:

- Reliability testing was conducted at the accountable-entity level:
  - The developer's approach to assessing accountable-entity, score-level reliability was to consider the extent to which assessments of a hospital using different but randomly selected subsets of patients produce similar measures of hospital performance. The developers refer to this as a split-sample or test-retest approach to reliability testing.
  - Measure score reliability, calculated using the split-sample approach, ranges from 0.230 for hospitals with at least two admissions to 0.628 for hospitals with at least 300 admissions.
  - The measure score reliability, as assessed by the split-sample method and with the updated minimum threshold of 50 cases, is 0.402. The developer states that because of the context in which the measure is used (pay-for-reporting program), this level of reliability is sufficient.

Validity

Preliminary ratings for validity: The SMP Passed on Validity with a score of: H-4; M-5; L-2; I-0

Validity Testing

- Validity testing was conducted at the accountable-entity level:
  - Empirical validity testing
    - The developer assessed the correlation of this new measure (AMI EDAC) with the existing CMS 30-Day AMI Readmission measure and calculated a Pearson correlation of 0.610 ($p < 0.0001$).
    - The correlation between AMI EDAC scores and the Star Rating readmission group score is -0.313 ($p<0.0001$), which suggests that hospitals with lower AMI EDAC scores (better performance) are more likely to have higher Star Rating readmission group scores (better performance).
    - The correlation between AMI EDAC scores and the Star Rating summary score is -0.221 ($p<0.0001$), which suggests that hospitals with lower AMI EDAC scores (better performance) are more likely to have higher Star Rating summary scores (better performance).
    - The correlation between AMI EDAC scores and AMI risk-standardized readmission rates (RSRRs) is 0.425 ($p<0.0001$), which suggests that hospitals with lower AMI EDAC scores (better performance) are more likely to have lower AMI RSRRs (better performance).
    - The developer also examined associations between AMI EDAC and components of Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) (in Table 4 of the measure information form). The developer found a negative, significant correlation with each of the components.

- The previous submission included validity testing using face validity conducted by a technical expert panel (TEP):
  - The developer assessed face validity using their 16 member TEP, including patient representatives, expert clinicians, researchers, providers, and purchasers.
  - Of the 16 members, 11 selected the "Agree" categories with the validity statement, or 91.7 percent.

Exclusions

- The measure includes four exclusion categories with a small proportion of patients excluded from the measure as specified:
  - Exclusion 1 (patients without at least 30 days of post-discharge enrollment in FFS Medicare for index admissions) accounts for 0.74 percent of all index admissions excluded from the initial cohort.
  - Exclusion 2 (patients who are discharged AMA) accounts for 0.66 percent of all index admissions excluded from the initial index cohort.
  - Exclusion 3 (patients with admission within 30 days of a prior index admission) accounts for 1.44 percent of all index admissions excluded from the initial index cohort.
  - Exclusion 4 (same-day discharges) accounts for 0.47 percent of the cohort.

Risk Adjustment

- The measure employs a hierarchical generalized linear model (HGLM) that consists of two parts, a logit model and a truncated Poisson model. The two-part logit/Poisson model (often called a "hurdle" model) assumes that the outcome results from two related processes: an initial dichotomous event (i.e., a patient has at least one acute care event), which is modeled as the logit of the probability of the event, and for patients with an event (those which clear the "hurdle"), the number of days, which is modeled as a Poisson process. The outcome, number of days, is a half-integer count variable (because ED visits count as 0.5 days).
- The random effects hurdle model has a c-statistic (Logistic model) of 0.6, and a deviance R-squared (Poisson model) of 0.061 (6.1 percent).
- The measure includes a statistical risk model with 31 risk factors.
- The developer has conducted extensive analysis on the conceptual rationale for social risk on this measure and empirical testing. The analyses show that patients with any of the three social risk factors (i.e., dual eligibility, low Agency for Healthcare Research and Quality [AHRQ] Socioeconomic Status [SES] Index, or race [Black]) are at increased risk of EDAC, even after adjusting for other risk factors in a multivariable model.
- The developer notes that the changes in measure scores between the adjusted and unadjusted measures are small, and measure scores estimated for hospitals with and without either social risk factor are highly correlated.
- The developer's decision regarding adjustment for social risk factors was based on the empiric results (i.e., the impact on model and measure scores), the conceptual model (e.g., hospitals are better able to mitigate the influence of social risk factors on the measured outcome than clinicians), and the use of the measure (in a payment program or for public reporting).

Meaningful Differences

- The developer demonstrated that out of 4,074 hospitals in the United States (U.S.), 219 had "fewer days in acute care than the U.S. national average," 1,157 had "no different from the U.S. national average," and 429 had "more days in acute care than the U.S. national average."

- A total of 2,269 hospitals were classified as "number of cases too small" (fewer than 50) to reliably tell how well the hospital is performing

Missing Data

- There were no missing data in the claims-based development and testing data.

Comparability

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

SMP Concerns

- SMP members noted some concerns around the specifications of the measure. One member noted that the numerator statement should contain the same criteria as the denominator and that the denominator should exclude patients that die during the index admissions. Another member noted that the rationale for setting an emergency department visit at 0.5 acute days is not specified and that if the measure is based on a three-year pooled sample, this should be noted in the specifications.
- Regarding reliability testing, the SMP was largely concerned about the measure's low reliability (0.402).
- The SMP did not have any major concerns regarding the validity testing, but one member did note that although the correlations were significant and in the expected direction, they were relatively weak.
- Regarding the risk adjustment approach, two members voiced concerns about model's predictive ability. One noted that the model may have limited ability to predict EDAC greater than zero or days of EDAC given EDAC greater than zero; coupled with low reliability this suggests that the measure has a substantial random component. Another SMP member suggested that the low c-statistic raises concerns about the validity of the outcome. Finally, one member found it unclear whether the clinical risk adjustors are present at index admission or discharge.

Subgroup 2

NQF #2789 Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care

Reason for withdrawal: Developer will assess feedback by the SMP and incorporate that into new, additional testing

Maintenance Measure

Brief Description of Measure: The Adolescent Assessment of Preparation for Transition (ADAPT) to Adult-Focused Health Care measures the quality of preparation for transition from pediatric-focused to adult-focused health care as reported in a survey completed by youth ages 16-17 years old with a chronic health condition. The ADAPT survey generates measures for each of the 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning.

Numerator Statement: The ADAPT survey consists of 26 questions assessing the quality of health care transition preparation for youth with chronic health conditions, based on youth report of whether specific recommended processes of care were received. The ADAPT survey generates measures for each of 3 domains: 1) Counseling on Transition Self-Management, 2) Counseling on Prescription Medication, and 3) Transfer Planning. ADAPT measure scores are calculated using the sum of the proportions of

positive responses to between 3 and 5 individual items. Complete instructions for measure score calculations are provided in the Detailed Measure Specifications (Appendix A).

1) Counseling on Transition Self-Management:

The numerator is the sum of the proportions of positive responses to the five questions about counseling on transition self-management, among respondents with valid responses to all questions.

2) Counseling on prescription medication:

The numerator is the sum of the proportions of positive responses to the three questions about counseling on prescription medication, among respondents who indicate that they take prescription medication every day and with valid responses to all questions.

3) Transfer planning:

The numerator is the sum of the proportions of positive responses to the four questions about transfer planning, among respondents who report being treated by a pediatric provider and with valid responses to all questions.

Denominator Statement: The target population of the survey is 16- or 17-year-old adolescents with a chronic health condition who are either (a) receiving health care services in a clinical program (for example, a sub-specialty practice focusing on management of a chronic condition, or a medical practice providing primary or preventative care, as opposed to urgent care) or (b) enrolled in a health plan or similar defined population.

The denominator for each measure is the number of respondents with valid responses for all of the questions in the measure.

Denominator Exclusions: SURVEY SAMPLE

Exclude patients in the following categories from the ADAPT survey sample frame:

1. "No-publicity" patients (i.e., those who requested that they not be contacted)
2. Court/law enforcement patients
3. Patients with a foreign home address
4. Patients who cannot be surveyed because of local, state, or federal regulations

SURVEY RESPONSE

Exclude survey respondents based on the following clinical and non-clinical criteria:

1. Undeliverable survey, i.e., the survey is returned by US Mail as undeliverable. "Undeliverable" should not be assumed merely because of non-response.
2. The survey is returned with clear indication that the patient does not meet eligibility criteria (e.g., ineligible age or lack of a chronic health condition).
3. Patient unable to complete survey independently: This must be indicated by the appropriate checkbox in the cover letter or equivalent clear indication by the parent/guardian that the patient is unable to complete the survey independently (e.g., due to cognitive limitation).
4. Exclude all respondents who answered "None" to ADAPT question 3 ("In the last 12 months, how many times did you visit this provider?").

Measure Type: Outcome: PRO-PM

Data Source: Instrument-Based Data, ADAPT Survey, ADAPT National Field Test Data Set

Level of Analysis: Clinician: Group/Practice, Facility, Health Plan

Risk-Adjusted: Statistical risk model with two factors

Sampling Allowed: Yes

Reliability

Preliminary ratings for reliability: The SMP Did Not Pass on Reliability with a score of: H-0; M-2; L-2; I-5

Specifications:

- The measure includes three measures of quality of healthcare transition preparation for youth with chronic health conditions: (1) Counseling on Transition Self-Management, (2) Counseling on Prescription Medication, and (3) Transfer Planning. Each of the three measures identifies specific questions from the ADAPT survey for the measures' calculation.
- The developer does not indicate that there is a roll-up score of the three measures. Instead, the developer describes that each of the three measures sums up the proportion of question responses with a value of one (indicating good quality of care) and divides this by the total number of respondents with valid responses, multiplying for a percentage.
- Measure specifications are clear and precise.
- Measure specifications have not changed since the last review.
- Measure specifications for the instrument-based measure also include the specific instrument (i.e., the ADAPT survey); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and the calculation of response rates to be reported with the performance measure results.
    - The instrument is available in English and Spanish.
    - Proxy responses are not allowed.
    - The measure includes instructions for creating a sampling frame and for collecting responses.

Reliability Testing:

- Reliability testing was conducted at the patient/encounter level:
    - It is unclear whether the developer has intended to present reliability testing at the patient/encounter level. The developer refers to previously submitted testing, but it is not presented here.
- Reliability testing was conducted at the accountable-entity level:
    - Typically, tests of internal consistency are reserved for items in a multi-item scale and tests at the patient/encounter level. However, the developer presents ordinal alpha results to test the consistency of three multi-item measures built from various questions in the ADAPT survey. The developer appears to be presenting these data to test the reliability of the measure scores (i.e., at the accountable-entity level).
    - The developer presents ordinal alpha results at the hospital, health plan, and clinic levels.
        - Hospital: The Counseling on the Transition Self-Management measure results in an ordinal alpha of 0.79, the Counseling on Prescription Medication measure

results in an ordinal alpha of 0.57, and the Transfer Planning measure results in an ordinal alpha of 0.99.

- Health Plans 1 and 2: The Counseling on Transition Self-Management measure results in an ordinal alpha ranging from 0.7–0.78, the Counseling on Prescription Medication measure results in an ordinal alpha ranging from 0.74–0.78, and the Transfer Planning measure results in an ordinal alpha of 0.99.
- Clinics #1, #2 and #3: The Counseling on Transition Self-Management measure results in an ordinal alpha ranging from 0.76–0.86, the Counseling on Prescription Medication measure results in an ordinal alpha ranging from 0.82–0.98, and the Transfer Planning measure results in an ordinal alpha ranging from 0.96–0.99.

- The developer reports that "sites in our field testing varied in their geographic location and demographic characteristics" and concludes that the internal consistency of the measures demonstrates sufficient reliability.

Validity

Preliminary ratings for validity: The SMP Did Not Pass on Validity with a score of: H-0; M-1; L-4; I-4

Validity Testing

- Validity testing was conducted at the patient/encounter level:
  - The developer performs a confirmatory factor analysis to test whether "the questions associated with each construct actually elicit information about the given construct." This testing is performed at the question level and so appears to support patient/encounter-level validity testing. No testing is presented for the third measure (Transfer Planning) because the developer did not have adequate sample sizes for testing.
  - Hospital-level results:
    - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor loading estimates ranged from 0.516–0.655. Standard errors ranged from 0.09–0.13. Two-tailed T-test results ranged from 5.027–6.615. The P-values were all less than 0.001.
    - Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.165–0.826. Standard errors ranged from 0.108–0.16. Two-tailed T-test results ranged from 1.527–5.163. The P-values ranged from less than 0.001 to 0.127.
  - Health Plan #1 results:
    - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor-loading estimates ranged from 0.332–0.694. Standard errors ranged from 0.075–0.114. Two-tailed T-test results ranged from 4.442–7.489. The P-values were all less than 0.001.
    - Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.576–0.673. Standard errors ranged from 0.08–0.089. Two-tailed T-test results ranged from 6.471–7.968. The P-values were all less than 0.001.
  - Health Plan #2 results:
    - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor-loading estimates ranged from

0.447–0.753. Standard errors ranged from 0.0092–0.113. Two-tailed T-test results ranged from 4.152–7.515. P-values were all less than 0.001.

- Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.408–0.643. Standard errors ranged from 0.11–0.119. Two-tailed T-test results ranged from 3.428–5.851. The P-values were all less than or equal to 0.001.

- Clinic #1 results:
  - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor-loading estimates ranged from 0.334–0.762. Standard errors ranged from 0.132–0.273. Two-tailed T-test results ranged from 1.57–5.674. The P-values ranged from 0.001–0.116.
  - Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.518–0.823. Standard errors ranged from 0.145–0.211. Two-tailed T-test results ranged from 3.568–5.09. The P-values were all less than 0.001.
- Clinic #2 results:
  - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor-loading estimates ranged from 0.472–0.864. Standard errors ranged from 0.135–0.353. Two-tailed T-test results ranged from 2.444–5.376. The P-values ranged from 0.001–0.015.
  - Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.229–1.15. Standard errors ranged from 0.103–0.508. Two-tailed T-test results ranged from 1.915–2.262. The P-values ranged from 0.024–0.056.
- Clinic #3 results:
  - Questions #4, #5, #6/7, and #8 relate to the construct of Counseling on Transition Self-Management measure. Factor-loading estimates ranged from 0.308–1.017. Standard errors ranged from 0.084–0.182. Two-tailed T-test results ranged from 1.699–12.124. The P-values ranged from less than 0.001 to 0.089.
  - Questions #10, #12, and #13 relate to the construct of Counseling on Prescription Medication measure. Factor-loading estimates ranged from 0.413–0.594. Standard errors ranged from 0.189–0.249. Two-tailed T-test results ranged from 2.047–2.841. P-values ranged from 0.004–0.041.

- Validity testing was conducted at the accountable-entity level:
  - The developer reports that "a 'gold standard' does not exist for determining the criterion validity of patient-reported measures of quality."
  - The developer presents a chi-square test of fit p-value, a Root Mean Squared Error of Approximation (RMSEA) (90% CI), Comparative Fit Index (CFI), and the Tucker Lewis Index (TLI) to test for goodness of fit for validating CFA models.
    - Hospital: Chi-square test of fit p-value = 0.013, RMSEA (90% CI) = 0.064 (0.028, 0.098), CFI = 0.892, and TLI = 0.826.

- Health Plans #1 and #2: Chi-square test of fit p-value ranged from less than 0.001 to 0.244, RMSEA (90% CI) ranged from 0.081 (0.061, 0.103) to 0.026 (0, 0.062), CFI ranged from 0.792 to 0.974, and TLI ranged from 0.664 to 0.958.
- Clinics #1, #2 and #3: Chi-square test of fit p-value ranged from 0.004 to 0.033, RMSEA (90% CI) ranged from 0.087 (0.026, 0.141) to 0.124 (0.069, 0.178), CFI ranged from 0.566 to 0.903, and TLI ranged from 0.393 to 0.864.

Exclusions

- The measure excludes the following groups of patients: those who requested to not be contacted; court/law enforcement patients; patients with a foreign home address; and patients who cannot be surveyed because of local, state, or federal regulations.
- The developer does not provide testing of the exclusions.

Risk Adjustment

- The measure uses a statistical risk model to adjust for two risk factors: respondent age and self-reported health status.
- The measure developer attests that there have been no changes to or updated testing performed on the risk adjustment model.

Meaningful Differences

- The developer performed t-tests and f-tests to compare the three measures' scores on case-mix adjusted model estimates. The t-test compared results between the two health plan sites. The f-test compared three sites (the two health plans and one hospital).
  - The Case-Mix Adjusted Measure Scores ranged from a low of 4 in Health Plan #1 to a high of 489 in Health Plan #2.
  - Results of the t-test: Counseling on Transition Self-Management = 0.028, Counseling on Prescription Medication = 0.267, and Transfer Planning = 0.225.
  - Results of the f-test: Counseling on Transition Self-Management = 0.024, Counseling on Prescription Medication = 0.075, and Transfer Planning = 0.158.
- The developer reports that the low results are a result of low scores overall across all sites.

Missing Data

- The developer describes how screening questions are used to determine whether questions are truly relevant to respondents to determine whether a response is missing or whether a question is appropriately skipped.
- The developer presents frequencies of truly missing responses for Hospital #1 and for Health Plans #1 and #2 for each question used to construct the three measures. Frequencies range from one to 11 and less than 3 percent of cases were truly missing. The developer reports the following: "The mean percentage missing was 1.3% and ranged from 0.67% for scheduling own appointments to 2.00% for discussing whether there is a need to change to a new provider who treats mostly adults."
- The developer compared respondents and non-respondents' demographic characteristics and medical complexity characteristics. The developer found small differences but reported that generally, the populations were similar. The developers' more detailed analysis showed a "higher proportion of 17 year-old adolescents in the Health Plan 1 respondent sample only" and "lower proportions of black patients in the respondent samples." The developer also found higher response rates for females in the hospital sample but not in the two health plans. The developer reported no differences based on chronic condition complexity.

- The developer compared respondents and non-respondents' demographic characteristics and medical complexity characteristics. The developer found small differences but reported that generally, the populations were similar. The developers' more detailed analysis showed a "higher proportion of 17 year-old adolescents in the Health Plan 1 respondent sample only" and "lower proportions of black patients in the respondent samples." The developer also found higher response rates for females in the hospital sample but not in the two health plans. The developer reported no differences based on chronic condition complexity.

Comparability

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

SMP Concerns

- A number of SMP members were concerned that ICD-9 coding was used in the specifications when NQF requires ICD-10 codes be used. One member requested more clarity around how entities are scored when domain scores differ within accountable entities. Another member noted that the specifications for risk adjustment may contribute more complication, and therefore more noise than signal, to the measure.
- There were mixed opinions regarding the reliability testing. Several SMP members were concerned that no patient/encounter level testing was presented, while others noted that it was presented but accountable entity level testing was not presented. The developer's submission was not clear to all reviewers regarding the type of testing presented. As many members noted, both levels are required for instrument based measures, such as PRO-PMs. One member was also concerned about the dates of data used for testing, stating that they were not updated since 2013-2014. This relates to the concern noted above that it is unclear how scores are calculated when domain scores differ within accountable entities. Another member also raised concern that the representativeness of the samples was never explained, raising doubt in the results with so few entities tested.
- The developers performed CFA to demonstrate the measures' validity at the patient/encounter level but some reviewers felt the results were too low as they were "below the acceptable level of 0.5 to support the uni-dimensionality of a domain." Additionally, the Transfer Planning domain was not tested.
- The reviewers were also largely concerned about the lack of accountable entity level validity testing presented. This is also required for instrument-based measures.
- The SMP members noted that although some analysis of meaningful differences was submitted, there was not a sufficient number of accountable entities sampled to adequately demonstrate those differences.
- Several SMP members were concerned about testing results to demonstrate validity of the risk adjustment model. Namely, no calibration statistics, decile plots, or calibration curves were provided. There were also concerns raised about the conceptual rationale for inclusion of age and health status, which were ultimately included in the model.