NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel – Measure Evaluation Web Meeting

The National Quality Forum (NQF) convened the Scientific Methods Panel (SMP) for a web meeting on October 25, 2022, for a discussion of the scientific acceptability (i.e., reliability and validity) of several complex measures submitted during the fall 2022 evaluation cycle. Of the 13 measures that the SMP reviewed during this cycle, five measures were discussed during the meeting, which included those that: (1) did not receive consensus from the SMP subgroups[1] and/or for which the SMP subgroups identified major areas of concern during their preliminary evaluations; (2) did not pass the SMP's preliminary evaluation but for which the measure developers provided additional information; or (3) were pulled for discussion by SMP subgroup members or NQF staff. This meeting summary includes brief summaries of the five measures and the final voting results taken during the meeting, the overarching methodological issues discussed during the web meeting, and the voting results for the measures that passed the SMP's preliminary evaluation and therefore were not discussed during the web meeting.

Welcome, Review of Meeting Objectives, Introductions, and Overview of Evaluation and Voting Process

Dr. Matthew Pickering, NQF senior director, welcomed the SMP members and participants to the web meeting. NQF staff reviewed the meeting objectives. The SMP members each introduced themselves and disclosed any conflicts of interest. Eric Weinhandl disclosed a conflict with NQF #3722 and NQF #3725; Sherrie Kaplan disclosed a conflict with NQF #2962; Susan White disclosed a conflict with NQF #3718, NQF #3720, and NQF #3721; and Zhenqiu Lin disclosed a conflict with NQF #3703. Eric Weinhandl was recused from NQF #3722 and NQF #3725 due to directly collaborating with the measure developer, with or without compensation. Sherrie Kaplan was recused from NQF #2962 due to working directly with a measure developer as a consultant. Susan White was recused from NQF #3718, NQF #3720, and NQF #3721 due to directly collaborating with the measure developer, with or without compensation. Zhenqiu Lin was recused from NQF #3703 due to their employment with the same organization that developed the measure. Following the introductions, Dr. Pickering and Hannah Ingber, NQF manager, reviewed the process for the measure discussions and the measure evaluation criteria.

Some SMP members were unable to attend the entire meeting due to early departures and late arrivals. However, the required attendance of 12 SMP members to hold the meeting was met and maintained for all measures for the entirety of the meeting. The vote totals reflect members present and eligible to vote. The quorum for voting in the two subgroups was eight SMP members, which was also maintained for all measures for the entirety of the meeting. The voting results are provided below.

Measure Evaluation

Forty-seven measures were submitted by the Intent to Submit deadline for the fall 2022 cycle. Thirteen measures were deemed as complex, and the SMP reviewed them for scientific acceptability. Each SMP

[1] Subgroups are determined by NQF staff. NQF assigns measures to subgroup members for evaluation based on panelists' relevant expertise, interests, measure-specific disclosures of interest, and Standing Committee membership.

member was assigned to one of two subgroups, and both subgroups were assigned either six or seven of the 13 measures being reviewed this cycle. The subgroup members then performed in-depth reviews and analyses of their assigned measures. The developers received these preliminary analyses prior to the web meeting and were given an opportunity to submit written responses to concerns expressed by the SMP reviewers. These responses were provided to the SMP prior to the meeting to prepare for its discussion and subsequent voting conducted by the subgroups on the measures during the meeting. Following the preliminary evaluation conducted by the SMP, two measures were withdrawn by the measure developer from the cycle, NQF #2881 and NQF #2789, leaving a total of 11 measures described in this summary. Measure evaluations that were discussed during the meeting were based on the preliminary analyses performed by the assigned SMP members.

During the meeting, the SMP evaluated the reliability, validity, and/or composite construction for five of the 11 measures due to the SMP's preliminary analysis results and/or additional information that was submitted by the developer for consideration during the SMP web meeting. The remaining six measures were not discussed during the meeting because the SMP subgroup members passed them on both reliability and validity and were not otherwise pulled for discussion. For these six measures, the subgroups' preliminary analyses serve as the final SMP assessment of scientific acceptability for the Standing Committees' consideration. For each of the five measures discussed during the meeting, NQF staff described the measure, noted the preliminary evaluation ratings of the respective SMP subgroup, and highlighted the criterion (or criteria) for which there was a lack of consensus and/or major areas of concern. Drs. David Nerenz and Christie Teigland, the SMP co-chairs, facilitated the remainder of the discussion. Lead discussants from the respective SMP subgroup first summarized the primary concerns of the SMP subgroup members. Other subgroup members were then given the opportunity to provide additional comments. Next, the SMP co-chairs invited comments or additional questions from other SMP members, and the developers were then invited to respond. Following these concerns and comments, NQF staff invited the measure developers to provide brief responses to the concerns raised by the subgroup members and to summarize their written response(s), if provided. The subgroup members then voted on the measure, producing the final votes of the SMP for the relevant criteria. These votes reflect the final assessment of scientific acceptability conducted by the respective SMP subgroups.

NQF's current policy dictates that a measure passes on the scientific acceptability criteria (i.e., reliability, validity, and composite construction) when greater than 60 percent of eligible voting members select a passing vote option (i.e., High or Moderate) on reliability, validity, and composite construction. A measure does not pass the SMP's review when less than 40 percent of voting members select a passing vote option on reliability, validity, or composite construction. The SMP has not reached consensus on the measure if between and including 40 and 60 percent of eligible voting members select a passing vote option on reliability, validity, or composite construction.

Voting Legend:

- *Scientific Acceptability: Reliability, Validity, and Composite Construction:* H – High; M – Moderate; L – Low; I – Insufficient; NA – Not Applicable

Subgroup 1

NQF #3725 Home Dialysis Retention (Kidney Care Quality Alliance [KCQA])

Description: Percent of all new home dialysis patients in the measurement year for whom >=90 consecutive days of home dialysis was achieved; **Measure Type**: Outcome: Intermediate Clinical

Outcome; **Level of Analysis**: Facility; Other; **Setting of Care**: Ambulatory Care; Home Care; Outpatient Services; Post-Acute Care; **Data Source**: Claims; Electronic Health Data; Electronic Health Records

Measure Steward/Developer Representatives at the Meeting

- Kathy Lester
- Feifei Ye
- Daniel Gilbertson

Final SMP Votes

- **Reliability**: Total Votes-8; H-0; M-5; L-2; I-1 (5/8 – 62.5%, Pass)
- **Validity**: Total Votes-11; H-1; M-7; L-2; I-1 (8/11 – 72.7%, Pass)

In its preliminary analysis, the SMP did not reach consensus on reliability but did pass the measure on validity. The SMP discussed and re-voted on reliability. The primary issue for discussion was the possible overestimation of the mean reliability estimate of 0.604 due to low sample sizes. The SMP noted that the denominator is a limited sample size due to only including incident home dialysis patients, a small population. The developer responded by explaining that small sample sizes are the reality for home dialysis patients. In acknowledging that the mean reliability estimate may be overestimated, the developer attempted to increase reliability using multiple years of simulated data, which they justified as a logical approach as home dialysis rates have remained relatively consistent over time. In analyzing multiple years of data, the developer found higher reliability estimates when calculated using more years of data. The developer acknowledged that using the simulated data may overstate the reliability but still asserted that the lack of variation indicates that the reliability estimates would still be high. The developer also pointed out that the specifications denote a "measurement year" and explained that when the Centers for Medicare & Medicaid Services (CMS) adopt the measure, they can define what is considered a measurement year through their implementation policies and are not restricted to a single year of data. CMS will also systematically exclude facilities with fewer than 11 patients as per CMS' standard reporting policies. Due to the considerations regarding small sample size and analyses presented by the developer, a three-year rolling average would likely be their recommendation to CMS for the implementation of this measure, although the developer noted that this is not how the measure is currently specified.

The SMP members discussed the methods for testing multiple years of data as one member commented during the SMP's preliminary review that they had concerns with the method used. Because an unreliable estimate of the provider probability at the entity level is likely with the small denominators, the SMP member described another method of calculation. The developer provided a response in the discussion guide using the member's suggested formula. The SMP considered these tests and also agreed that because the measure is going to be implemented with multiple years of data, the reliability will be higher than originally presented. Additionally, because this is a new measure, it will return for endorsement maintenance within three years and can provide additional testing at that time. The SMP noted that this measure is specified as a one-year measure and that the measure reliability of 0.604 is acceptable using one year of data and that this should be moved to the Standing Committee for further discussion. The SMP re-voted to pass the measure on reliability.

NQF #3654 Hospice Care Index (Centers for Medicare & Medicaid Services [CMS]/Abt Associates)

Description: The Hospice Care Index monitors a broad set of leading, claims-based indicators of hospice care processes. The ten indicators reflect 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. Index scores are calculated as the total instances a hospice exceeds a threshold for

each of the 10 indicators. The index thereby seeks to identify hospices which are outliers across an array of multifaceted indicators, simultaneously.; **Measure Type**: Composite; **Level of Analysis**: Facility; **Setting of Care**: Behavioral Health; Inpatient/Hospital; Other; Outpatient Services; Post-Acute Care; Home Care; **Data Source**: Claims

Measure Steward/Developer Representatives at the Meeting

- T.J. Christian
- Zinnia Harrison
- Rebekah Natanov
- Ihsan Abdur-Rahman

Final SMP Votes

- **Reliability**: Total Votes-11; H-1; M-2; L-3; I-5 (3/11 – 27.3%, No Pass)
- **Validity**: Total Votes-11; H-0; M-3; L-2; I-6 (3/11 – 27.3%, No Pass)
- **Composite Quality Construct:** Total Votes-10; H-1; M-4; L-3; I-2 (5/10 – 50.0%, Consensus Not Reached [CNR])

In its preliminary analysis, the SMP did not pass the measure on reliability or validity and did not reach consensus on the composite quality construct. For reliability, the SMP noted concerns with the insufficiency of the stability analysis presented. No empirical test that provides information about the measurement error was conducted. Moreover, SMP subgroup members expressed concern about the lack of reliability testing for individual items within the composite. The developer responded by explaining that traditional signal-to-noise testing was not applicable given the nature of the composite and did not provide any additional testing data. The SMP suggested that the developer consider employing a test-retest approach or split-sample testing as well as testing the reliability of the individual items within the composite.

For validity, the SMP identified concerns with the low Pearson correlations in the validity testing results and lack of risk adjustment of the overall composite and of individual items within the composite that signify outcomes. The developer responded by explaining that because the composite focuses on outliers, risk adjustment likely would not impact a hospice's scoring, except for those hospices at the threshold. Regarding the composite construction, the SMP, in addition to noting earlier concerns with reliability and validity testing, noted that the information was insufficient to assess the construction of the composite. The SMP suggested the developer consider providing a logic model, conducting correlational analysis if it is designed to be a reflective composite, and providing stronger empirical support for including each component in the composite. After discussing the measure, the SMP did not re-vote on reliability, validity, or the composite construction. Therefore, the SMP did not pass the measure on reliability or validity and did not reach consensus on the composite construction.

This measure is not eligible for a revote from the Geriatrics and Palliative Care Standing Committee because the SMP determined that inappropriate methods were used, namely, no appropriate empirical tests of reliability were conducted. In addition, the individual items within the composite were not tested for reliability and validity.

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 New Haven Health Services Corporation – Center for Outcomes Research and Evaluation [CORE])

Description: 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. 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; **Measure Type**: Composite; **Level of Analysis**: Health Plan; Population: Regional and State; **Setting of Care**: Ambulatory Care; **Data Source**: Claims; Other (specify)

Final SMP Votes

- **Reliability**: Total Votes-10; H-5; M-5; L-0; I-0 (10/10 – 100%, Pass)
- **Validity**: Total Votes-10; H-1; M-8; L-1; I-0 (9/10 – 90.0%, Pass)
- **Composite Quality Construct:** Total Votes-10; H-2; M-6; L-1; I-1 (8/10 – 80.0%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure during the fall 2022 cycle.

NQF #2651 CAHPS® Hospice Survey, Version 9.0 (CMS)

Description: The measures submitted here are derived from the CAHPS® Hospice Survey, which is a 47-item standardized questionnaire and data collection methodology. The survey is intended to measure the care experiences of hospice patients and their primary caregivers. Respondents to the survey are the primary informal caregivers of patients who died under hospice care. The hospice identifies the primary informal caregiver from their administrative records. Data collection for sampled decedents/caregivers is initiated two months following the month of the decedent's death. The publicly reported measures described here include the following six multi-item measures: 1) Hospice Team Communication, 2) Getting Timely Care, 3) Treating Family Member with Respect, 4) Getting Emotional and Religious Support, 5) Getting Help for Symptoms, and 6) Getting Hospice Training. In addition, there are two global rating items that are publicly-reported measures: Rating of the hospice care and Willingness to recommend the hospice; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Facility; **Setting of Care**: Other; Inpatient/Hospital; Home Care; **Data Source**: Instrument-Based Data

Final SMP Votes

- **Reliability**: Total Votes-11; H-6; M-3; L-2; I-0 (9/11 – 81.8%, Pass)
- **Validity**: Total Votes-11; H-1; M-6; L-2; I-2 (7/11 – 63.6%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The Geriatrics and Palliative Care Standing Committee will evaluate this measure during the fall 2022 cycle.

NQF #3726 Serious Illness Survey for Home-Based Programs (RAND Corporation)

Description: 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.

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.; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Other; **Setting of Care**: Home Care; **Data Source**: Instrument-Based Data

Final SMP Votes

- **Reliability**: Total Votes-11; H-4; M-4; L-2; I-1 (8/11 – 72.7%, Pass)
- **Validity**: Total Votes-11; H-3; M-6; L-2; I-0 (9/11 – 81.8%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The Geriatrics and Palliative Care Standing Committee will evaluate this measure during the fall 2022 cycle.

NQF #3722 Home Dialysis Rate (KCQA)

Description: Percent of all dialysis patient-months in the measurement year in which the patient was dialyzing via a home dialysis modality.; **Measure Type**: Process; **Level of Analysis**: Facility; Other; **Setting of Care**: Ambulatory Care; Home Care; Outpatient Services; Post-Acute Care; **Data Source**: Claims; Electronic Health Data; Electronic Health Records

Final SMP Votes

- **Reliability**: Total Votes-11; H-5; M-3; L-1; I-2 (8/11 – 72.7%, Pass)
- **Validity**: Total Votes-11; H-1; M-6; L-3; I-1 (7/11 – 63.6%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The Renal Standing Committee will evaluate this measure during the fall 2022 cycle.

Subgroup 2

NQF #3721 Patient-Reported Overall Physical Health Following Chemotherapy Among Adults With Breast Cancer (Purchaser Business Group on Health)

Description: 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.; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Clinician: Group/Practice; **Setting of Care**: Ambulatory Care; Outpatient Services; **Data Source**: Electronic Health Records; Instrument-Based Data; Paper Medical Records

Measure Steward/Developer Representatives at the Meeting

- Rachel Brodie
- Kristen McNiff Landrum
- FeiFei Ye

Final SMP Votes

- **Reliability**: Total Votes-10; H-0; M-2; L-8; I-0 (2/10 – 20.0%, No Pass)
- **Validity**: Total Votes-10; H-0; M-2; L-5; I-3 (2/10 – 20.0%, No Pass)

The SMP's discussion of NQF #3721 focused on reliability and validity, during which the preliminary vote of the SMP subgroup members had been "no pass" and "consensus not reached" (CNR), respectively. During the meeting, the developer noted that they did not submit any additional reliability testing but instead recontextualized their testing results. The developer explained that when a threshold of 0.6, rather than 0.7, was used, it reduced the minimum number of patients to achieve reliability to 43 patients, whereas with the 0.7 threshold, the minimum sample size was 66 patients. In addition, when 0.6 was used, 50 percent of groups reached this threshold, whereas with the 0.7 threshold, 10 percent of groups reached this threshold.

One SMP member stated that this measure was both qualitatively and quantitatively different than the other two similar measures, NQF #3720 and NQF #3718, both of which are described below. The SMP noted that this measure had poorer internal consistency compared to the other two measures. The developer acknowledged that the measure had lower reliability possibly because the between-group variation for physical health is smaller than for the other two measures. The developer also stated that an additional analysis has not been done to better understand this but noted that since the signal is smaller, the signal-to-noise ratio (i.e., the intraclass correlation coefficient [ICC]) might be lower.

The SMP suggested that larger groups should be tested since the average group size actually tested was 32, which is lower than the number of patients required to be reliable (i.e., 66). The developer stated that this would occur in the future. A question was raised regarding whether there was a reason the sample size was limited. The developer explained that it was related to the coronavirus disease 2019 (COVID-19) pandemic, which limited the ability to test the measure. The SMP again noted that the solution to the limited sample size issue is to gather more data, which would also likely help address the issue with meaningful differences, which was part of the validity testing. Due to these continued concerns, the SMP did not request to re-vote on reliability.

Regarding validity, the SMP raised several concerns. The first issue involved the face validity testing, which consisted of eight votes that the measure can differentiate good from poor quality among accountable entities and four technical expert panel (TEP) members who did not vote based on the testing implications and COVID-19 impacts. The developer stated that the TEP members were concerned with how the measure may have been impacted by COVID-19, specifically the small sample sizes and performance scores. It was noted that empirical validity testing was performed with this measure and other related measures in the additional information included in the discussion guide, in which correlations were moderate in the correct direction.

With respect to the risk adjustment model, there was also incomplete data for some risk factors. In the updated risk adjustment results presented by the developer in the discussion guide, only three of the 13 variables were statistically significant due to the limited sample size. The developer noted that in moving forward with testing, these variables should still be candidates for possible inclusion in risk adjustment models with future testing because they were chosen by the TEP.

There were also concerns about missing data and how the developer handled the missingness in their measure calculation. The developer responded by explaining that the missing data were imputed for comorbidities. There were also concerns regarding nonresponse bias. An SMP member noted that the response rate was 37 percent, which is low. Therefore, the ability to impute is limited, especially with such a low sample size. The developer also responded to concerns about nonresponse bias and presented a paper from the literature that supported their approach.

There were also minimal meaningful differences between clinician groups, with only one of 10 groups being statistically different (one higher than average). The last issue addressed the timing of the baseline survey, which was conducted between two weeks before and one week after oral chemotherapy. SMP members questioned whether the baseline survey was conducted after the oral chemotherapy started and whether it could potentially impact the baseline survey score. The developer clarified that a meaningful difference on a physical health score in cancer patients from the literature was between a three- to six-point difference in a T-score scale of 50. It was noted that the group scores that were above the average equaled 5.19 points, which is more than half of the standard deviation. Regarding the timing of the baseline score, one SMP member, who is also a medical oncologist, stated that this could be a concern in which oral chemotherapy is more of a mainstay of treatment. However, in this population of breast cancer patients, chemotherapy is predominantly intravenous rather than oral, and therefore, it would not be a large concern with this particular measure. An SMP member stated that an overarching issue from the discussion was the low sample size; they expressed a desire to see the measure retested with more sites and a higher sample size. The SMP decided to re-vote based on the discussions on validity. With the revote, the SMP did not pass the measure on validity.

NQF #3720 Patient-Reported Fatigue Following Chemotherapy Among Adults With Breast Cancer (Purchaser Business Group on Health)

Description: 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.; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Clinician: Group/Practice; **Setting of Care**: Ambulatory Care; Outpatient Services; **Data Source**: Electronic Health Records; Instrument-Based Data; Paper Medical Records

Measure Steward/Developer Representatives at the Meeting

- Rachel Brodie
- FeiFei Ye
- Kristen McNiff Landrum

Final SMP Votes

- **Reliability**: Total Votes-10; H-0; M-9; L-1; I-0 (9/10 – 90.0%, Pass)
- **Validity**: Total Votes-10; H-0; M-6; L-3; I-1 (6/10 – 60.0%, CNR)

The preliminary vote for NQF #3720 was a pass for reliability and CNR on validity. The CNR decision for validity was due to concerns with face validity testing, the lack of demonstration of meaningful differences, and the missing response rates, as it had been with NQF #3721.

The SMP noted that this measure was similar to both NQF #3721 and NQF #3718. However, since the reliability testing results passed the SMP's preliminary review, the SMP only discussed the validity testing concerns.

An SMP member stated that there were similar concerns with this measure regarding validity as there were for NQF #3721, notably, the missing data and nonresponse bias, as well as the face validity. The results of the face validity vote did not demonstrate strong agreement among the TEP. Only three of eight TEP members agreed or strongly agreed that the measure could differentiate quality of care. One SMP member commented that given the way the scale was set up for face validity, in which a score of three out of five was "moderately agree," eight out of eight TEP members either strongly agreed, moderately agreed, or agreed that the measure could differentiate good versus poor quality of care. However, another SMP member commented that the four TEP members who did not vote stated that the pandemic confounded the issue of validity for this measure due to the inability of the measure to parse whether the fatigue is due to COVID-19 or cancer. Additionally, an SMP member stated that some of the issues in this measure were the same as the ones in the previous measure, NQF #3721 (specifically the small sample size; nonresponse; and missingness, considering it was the same data set). However, meaningful differences in performance were less of a concern for NQF #3720 because in this case, two out of the 10 clinician groups were statistically different (one higher and one lower). Many SMP members agreed that fewer threats to validity exist in NQF #3720 compared to NQF #3721.

The developer stated the rationale of the measure: Patients who undergo chemotherapy with curative intent tend to have long-term symptoms, including pain, fatigue, and trouble with physical and mental health. These issues can persist for months or years. The developer noted that there is also evidence to show that clinicians' care practices can prevent some of these symptoms in the survivorship phase of treatment. Regarding fatigue being more sensitive to the COVID-19 pandemic, the developer reminded the SMP that fatigue is also assessed at the baseline for the purposes of risk adjustment. Regarding the face validity votes, the developer stated that while only three TEP members strongly agreed, there were no votes of one or two out of five from the TEP that rated the face validity, which would have indicated disagreement had those votes been received.

An SMP member further questioned via comment the relationship between the COVID-19 pandemic, cancer care, and fatigue. COVID-19, in particular, can cause fatigue, particularly long-COVID, and confound the measure. However, another SMP member stated that this would be difficult to sort out methodologically, even after the pandemic ends and COVID-19 becomes endemic. Ultimately, the SMP re-voted on validity but still did not reach consensus on this criterion.

NQF #3718 Patient-Reported Pain Interference Following Chemotherapy Among Adults With Breast Cancer (Purchaser Business Group on Health)

Description: 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.; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Clinician: Group/Practice; **Setting of Care**: Ambulatory Care; Outpatient Services; **Data Source**: Electronic Health Data; Instrument-Based Data; Paper Medical Records; Electronic Health Records

Measure Steward/Developer Representatives at the Meeting

- Rachel Brodie
- FeiFei Ye
- Kristen McNiff Landrum

Final SMP Votes

- **Reliability**: Total Votes-10; H-0; M-9; L-1; I-0 (9/10 – 90.0%, Pass)
- **Validity**: Total Votes-10; H-2; M-5; L-1; I-2 (7/10 – 70.0%, Pass)

NQF staff pulled this measure for discussion on validity to ensure the criteria are applied consistently as it is grouped with NQF #3720 and NQF #3721. While the other two measures in the group preliminarily received CNR votes, NQF #3718 passed on both reliability and validity in the preliminary vote. An SMP member mentioned that it would be important to differentiate why the SMP did not reach consensus on NQF #3720 on validity, while the SMP recommended to pass NQF #3718 on validity. One SMP member stated that the Standing Committee would also question the CNR vote versus a passing vote on validity and would need to consider this in their evaluation. Another SMP member stated that in a side-by-side comparison of NQF #3720 and NQF #3718, they could not see any material reason why one would be CNR and why one would pass based on the objective validity testing results as well as the approach. By contrast, another SMP member mentioned that the face validity results were notably better in NQF #3718 than NQF #3720, which justified the difference. The SMP discussed and observed that the votes were not all that different, a 6–4 vote for NQF #3720 compared to a 7–3 vote for NQF #3718. The SMP chose not to re-vote on reliability or validity for NQF #3718 due to these differences in testing results, and therefore, the measure passed on both criteria.

NQF #2958 Informed, Patient Centered (IPC) Hip and Knee Replacement Surgery (Massachusetts General Hospital)

Description: The measure is derived from patient responses to the Hip or Knee Decision Quality Instruments. Participants who have a passing knowledge score (60% or higher) and a clear preference for surgery are considered to have met the criteria for an informed, patient-centered decision. The target population is adult patients who had a primary hip or knee replacement surgery for treatment of hip or knee osteoarthritis.; **Measure Type**: Outcome: PRO-PM; **Level of Analysis**: Clinician: Group/Practice; **Setting of Care**: Ambulatory Care; Outpatient Services; **Data Source**: Instrument-Based Data

Final SMP Votes

- **Reliability**: Total Votes-9; H-6; M-2; L-0; I-1 (8/9 – 88.9%, Pass)
- **Validity**: Total Votes-9; H-4; M-4; L-1; I-0 (8/9 – 88.9%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The Patient Experience and Function Standing Committee will evaluate this measure during the fall 2022 cycle.

NQF #2962 Shared Decision-Making Process (Massachusetts General Hospital)

Description: 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. This proposal is to focus on patients who have undergone any one of 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.; **Measure Type**: Outcome: PRO-

PM; **Level of Analysis**: Clinician: Group/Practice; **Setting of Care**: Ambulatory Care; Inpatient/Hospital; Outpatient Services; **Data Source**: Instrument-Based Data

Final SMP Votes

- **Reliability**: Total Votes-10; H-0; M-8; L-0; I-2 (8/10 – 80.0%, Pass)
- **Validity**: Total Votes-10; H-3; M-4; L-1; I-2 (7/10 – 70.0%, Pass)

In its preliminary analyses, the SMP passed the measure on both reliability and validity. Subgroup members also found the measure to be both reliable and valid. The Patient Experience and Function Standing Committee will evaluate this measure during the fall 2022 cycle.

Discussion of Overarching Methodological Issues Identified During Measure Evaluation

Evaluation of Face Validity

The issue of evaluating face validity has been a topic of conversation for the SMP for some time, specifically, grappling with determining an acceptable degree of agreement among members of the measure's TEP.

For three measures this cycle, NQF #3718, NQF #3720, and NQF #3721, face validity was heavily debated. In each of these measures, four of the 12 TEP members who assessed the measure's face validity declined to participate in voting due to concerns with how the measure may have been impacted by COVID-19, specifically the small sample sizes and performance scores, and requested additional testing. The SMP members appreciated the developer's explanation for the four TEP members' decision to abstain from voting because it helped them to decide on the face validity results. The members also encouraged the developer to continue to provide an explanation in their full measure submission.

The differentiation between "agreement" and "moderate agreement" in face validity assessment scales was a source of conversation for the SMP members and how the face validity results should be interpreted for NQF #3720. Some members posited that this distinction was important in showing the two measures' relative perceived abilities to distinguish good versus poor quality of care, while other SMP members proposed that the distinction between agreement and moderate agreement was not as important for their votes on validity.

The SMP also raised issue with the composition of the developer's TEP, noting that it is integral to include patient and caregiver representatives and perspectives. However, the SMP appreciated the developer's efforts to include these representatives in prior steps and other parts of the measure development process, such as selecting the Patient-Reported Outcomes Measurement Information System (PROMIS) scales for assessing patient-reported outcomes, participating in the patient-reported outcomes in oncology (PROMOnc) steering committee, and helping with the implementation of a patient burden questionnaire during testing.

The SMP encouraged NQF staff to consider how NQF criteria and guidance may be updated to help any SMP or Standing Committee discussions on the level of agreement needed to achieve and the TEP's composition for achieving sufficient face validity, given that as how it stands now, the NQF face validity criteria are not that prescriptive.

COVID-19 Impact on Measure Development

The SMP has discussed the long-term impacts that the COVID-19 pandemic would have on measure development, particularly issues with measure testing, such as sample size limitations and/or potential confounding.

During the fall 2022 measure evaluation meeting, it was noted for several of the measures that the sample sizes were limited due to the COVID-19 pandemic, leading to less certainty surrounding measure reliability and validity.

In addition to the small sample sizes, questions were raised for the three grouped cancer measures (NQF #3718, NQF #3720, and NQF #3721) as to how much COVID-19-related fatigue or pain could be confounding the measures. This is an important consideration because validity is testing whether the measure is being used correctly as an indicator of quality of care. In the case of these measures, it is not clear whether the measure is truly capturing cancer fatigue (or even pain) versus fatigue (or pain) brought on by COVID-19. Therefore, the SMP noted that it is difficult to parse out in these measures, in particular, what pain and/or fatigue may be related to COVID-19 versus being related to cancer. However, the SMP noted that moving forward, it may be necessary to consider including COVID-19 diagnosis as a factor in risk adjustment models to account for it as a possible confounder.

As measure development moves forward, the SMP encouraged measure developers to consider and continue to explain the impact COVID-19 may have on a measure's reliability and validity in their NQF submissions.

Public Comment

Dr. Pickering opened the lines for NQF member and public comments. One member of the public commented that the SMP has been discussing reliability thresholds and additional reliability and validity requirements for a long time and that they hope for their implementation in 2023 as they fit into a larger issue of reliability in the Consensus Development Process (CDP). The commenter continued by noting that a study referenced in a developer's submission was actually an abstract poster presented and only consisted of 198 patients with breast cancer; they further noted that the study was a non-randomized, quality improvement research study. The commenter stated that the study was limited and raised a continued concern: Evaluators need to ask the developer to assess the evidence they provide carefully.

Next Steps

Ms. Ingber reviewed the next steps and reminders for the SMP and the measures that the SMP members reviewed during this cycle. NQF staff will inform the developers and Standing Committees of the SMP's discussion and votes. Measures that passed on both reliability and validity or for which consensus was not reached will be considered by the relevant Standing Committees during the spring 2022 evaluation cycle.

According to NQF endorsement guidance, measures that did not pass the SMP's vote may be pulled for discussion by the relevant Standing Committee. However, measures are not eligible for a revote if any of the following are true: (1) An inappropriate methodology or testing approach was applied to demonstrate reliability or validity; (2) Incorrect calculations or formulas were used for testing; (3) A description of the testing approach, results, or data was insufficient for the SMP to apply the criteria; or (4) Appropriate levels of testing were not provided or otherwise did not meet NQF's minimum evaluation requirements.

NQF #3654 is ineligible for a revote due to the inappropriate methodology used to demonstrate reliability and/or validity. NQF #3721 is eligible for discussion and a revote from the Standing Committee if it is pulled by the Patient Experience and Function Standing Committee. Measures moving forward in the fall 2022 evaluation cycle will be reviewed by their respective Standing Committees and discussed by the Consensus Standards Approval Committee (CSAC).