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 26–27, 2021, to evaluate the fall 2021 cycle measures. A total of 15 complex measures were submitted for the fall 2021 cycle. However, three of those measures were withdrawn prior to the SMP's review. Of the 12 measures the SMP reviewed during the preliminary analyses, seven measures were discussed during the web meeting, including those that: (1) did not pass or did not reach consensus, (2) included major areas of methodological concern, and (3) included specific requests from SMP subgroup members or NQF staff based on the preliminary evaluation findings. This meeting summary includes brief summaries of the seven measures discussed and overarching methodological evaluation and process issues discussed during the web meeting. Of the seven SMP-discussed measures, three measures passed the SMP's evaluations, three did not pass and were ineligible to move forward for further evaluation, and one measure was withdrawn after the SMP's measure discussion. The meeting summary also includes the voting results for the five measures that passed the scientific acceptability evaluations within their respective SMP subgroup preliminary analyses that did not require discussion from the SMP during the web meeting.

## Welcome, Introductions, and Review of Meeting Objectives

Tricia Elliott, NQF senior managing director, welcomed the members of the SMP, measure developers, other NQF staff, and members of the public to the web meeting. Dana Gelb Safran, NQF president and CEO, and SMP Co-Chairs David Nerenz and Christie Teigland also provided opening remarks. Ms. Elliott asked the SMP members to introduce themselves and provide any disclosures or conflicts of interest relevant to the measures to be discussed during the meeting. A detailed list of recusals can be found in the fall 2021 SMP member recusals document. SMP members who were recused for the following measures included Dr. Patrick Romano, who identified a conflict of interest with measures #3633e, #3662e, and #3663e; Dr. Samuel Simon, who identified a conflict of interest with measures #3633e, #3639, #3649e, #3650e, #3652e, #3662e, and #3663e; and Dr. Zhenqiu Lin, who identified a conflict of interest with measures #3638, #3639, and #3667. These SMP members did not participate via the web meeting chat or verbally or via email during the measure discussions of the respective measures.

## Overview of Spring 2021 Evaluations and Evaluation Processes

NQF staff provided an overview of the 29 measures the SMP reviewed during spring 2021 measure cycle. Of those 29, 23 of them passed the SMP's evaluation and were considered for endorsement recommendation by their respective Consensus Development Process (CDP) Standing Committees. Two different CDP Standing Committees re-voted on the scientific acceptability criteria for two of the 29 measures the SMP reviewed. For both measures, the SMP did not reach consensus on validity, while the respective Standing Committees passed the measures. For the spring 2021 cycle, overall, there was an 86 percent agreement rate between the CDP Standing Committees and the SMP for the spring 2021 measures. One SMP member requested that future SMP and CDP Standing Committee voting metrics include the CDP Standing Committees' voting differences from the SMP, as well as those that accepted the scientific acceptability votes from the SMP. After reviewing the spring 2021 cycle evaluation results, NQF staff introduced the 12 measures under the SMP's review for the fall 2021 evaluation cycle.

Additionally, Ms. Elliott described the process for the measure discussions and voting and highlighted key NQF measure evaluation criteria specific to the SMP's reviews. This information included measure evaluation ratings, meeting and voting quorum, differences in testing requirements by measure type, testing and evaluation guidance, and eligibility criteria for respective CDP Standing Committees to evaluate measures that did not pass the SMP's evaluation.

## Measure Evaluations

During the meeting, the SMP evaluated 12 complex measures, including one maintenance and 11 new measures for scientific acceptability consideration. Complex measures are defined as outcome measures, including intermediate clinical outcomes, instrument-based measures (e.g., patient-reported outcome-based performance measures [PRO-PMs]), cost/resource use measures, efficiency measures (i.e., those combining concepts of resource use and quality), and composite measures. The measure evaluation ratings were based on the preliminary analyses performed by assigned SMP members who were assigned to one of two subgroups. Each subgroup member performed in-depth reviews and analyses of their six assigned measures. Developers received the summaries of the preliminary analyses prior to the web meeting and were provided an opportunity to submit written responses to concerns expressed by SMP reviewers. Developer responses were provided to the SMP members prior to the meeting to prepare for measure discussions and final measure voting by subgroups during the meeting. The preliminary analyses for each measure, including overviews, results, discussion items, and developer responses following the preliminary analyses, were provided in the SMP Fall 2021 Discussion Guide.

During the meeting, the SMP evaluated the applicable reliability and/or validity for seven measures based on their preliminary analyses and additional information submitted for consideration by the developers. For each measure discussed, NQF staff described the measure, provided the subgroup's preliminary evaluation ratings for each scientific acceptability criterion, noted the criterion (or criteria) for which the measure "*did not pass*" or for which "*consensus was not reached*," and/or outlined any applicable major areas of methodological concern identified by the SMP subgroup members or NQF staff. To ensure reliability is consistently evaluated across the subgroups, any measure with a reliability result of 0.50 or lower was pulled for the SMP's discussion. For this cycle, no measures met this criterion. Drs. Nerenz and Teigland facilitated the measure reviews. Lead discussants from the SMP subgroups summarized the primary concerns and asked targeted questions for subgroup consideration and feedback. Measure developers were invited to provide a brief measure introduction, offer responses to raised subgroup concerns, and summarize any written responses provided to the SMP following preliminary analyses. Next, the co-chairs invited comments or additional questions from other SMP members, and the developers provided additional responses as requested. After the measure discussion concluded for a discussed criterion, a quorum with a minimum of 66 percent of subgroup members (i.e., nine subgroup members for subgroup 1 and eight subgroup members for subgroup 2) was required and achieved for all measure votes. Subgroup votes conducted during the meeting produced the final SMP assessment of scientific acceptability criteria for the CDP Standing Committees' consideration. The remaining five measures evaluated by the SMP subgroups during the fall 2021 cycle passed all applicable scientific acceptability criteria, and therefore, they were not pulled for discussion during the meeting. For these measures, the subgroup's preliminary analyses serve as the final SMP assessment of scientific acceptability for the CDP Standing Committees' consideration.

The criterion voting options are listed below:

**Rating Scale:** H – High; M – Medium; L – Low; I – Insufficient; NA – Not Applicable

## Subgroup 1

During the meeting, subgroup 1 discussed six measures (#3649e, #3650e, #3652e, #3638, #3639, and #3667). This subgroup re-voted validity for #3667 and both reliability and validity for #3649e. The developer withdrew #3652e after several SMP members raised concerns with the risk adjustment model during the meeting. The final results for the six measures evaluated by subgroup 1 are presented below.

### #3649e Risk-Standardized Complication Rate (RSCR) Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) Electronic Clinical Quality Measure (eCQM) (Brigham and Women's Hospital [BWH])

#### *Measure Steward/Developer Representatives at the Meeting*

Patricia Dykes, Mica Bowen, Stuart Lipsitz

#### *Scientific Methods Panel Votes*

- **Reliability**: H-0; M-2; L-7; I-1 (No Pass)
- **Validity**: H-0; M-6; L-4; I-0 (Consensus Not Reached)

This new electronic clinical quality measure (eCQM) was pulled for the SMP's discussion, as consensus was not reached on the reliability criterion. The developer's submission presented the eFeasibility Scorecard to assess electronic health record (EHR) data availability, accuracy, terminology standards, and workflow with all 23 data elements scoring 100 percent in both Cerner (11 sites) and Massachusetts General and Brigham's (MGB) Epic sites (six sites). All data elements used to calculate performance were readily captured across MGB facilities and clinician groups. A few SMP members expressed their concern that only two EHRs were used for testing, stating that approximately 50 percent of all patient EHR documentation does not use the two enterprise systems. Subgroup members were concerned that data element agreement was only conducted in one EHR; however, the developer clarified that multiple Cerner EHRs were used in the testing. NQF staff confirmed that NQF's requirement for EHR data is "more than one"; therefore, it meets the requirement.

The developer conducted patient/encounter reliability testing using Epic EHR source mart data within the MGB enterprise data warehouse (EDW) and a split-half approach to test the reliability of the predicted/expected ratios at the clinician group level, comparing the agreement across clinician groups on the performance measure. Numerous social risk variables were analyzed for measure testing, including the African American race, smoking status, ZIP code (proxy for income level), English as a primary language, and body mass index (BMI). All data elements were pulled from the EHR, and ZIP codes were analyzed to United States (U.S.) Census data. The developer reported a Spearman rank correlation between the two split-half samples of r = 0.978. For variability across clinician groups, the intraclass correlation coefficient (ICC) equaled 0.006 (95 percent confidence interval [CI]: -0.017-0.027), which raised concerns related to the reliability of the measure in practice. The developer surmised that the very low ICC is likely due to the small clinician group sizes and small performance variation across groups. One SMP member raised concern with the generalizability of the results for 17 providers, while another member questioned the four years of data used for testing and validation that may contain higher sample sizes in each split-half sample than would be available in a single performance year. They also noted this sampling strategy may have affected the Spearman correlation coefficient results in overestimates for the two years of data. Another SMP member raised concerns with the wide CI surrounding the ICC estimates and the different versions of the ICC estimates in the testing results, along with the statistical uncertainty of the high correlations between point estimates across training and validation samples.

During the measure discussion, the lead discussant noted the small reliability ICC samples' sizes and the less than 1 percent performance variance between clinical groups. The lead discussant further stated that the results demonstrated greater "within" than "between" group variation, which may mask a skew when the measure shifts from hospital to group accountability. One SMP member stated that the variability between hospital complication rates is also attributable to differences in case mix and not to differences in provider performance. Multiple SMP members noted that additional insights could be gleaned with higher test volumes and that the volumes were insufficient to evaluate the results. The lead discussant also noted a very high exclusion rate of nearly 50 percent, thus resulting in a significantly reduced working denominator. Furthermore, the lead discussant was concerned with the results of the EHR data abstraction and significant population differences in osteoarthritis between EHRs, as this diagnosis should be very high in patients who have undergone total hip arthroplasty (THA) and total knee arthroplasty (TKA) surgeries. This SMP member noted the Cerner sites' data only included approximately 15 percent of patients with osteoarthritis. Other members questioned how to rate the reliability criterion with two divergent results (i.e., high split-half and low ICC). As an accountability measure, would the ICC be more indictive of reliability? Following the discussion, the subgroup agreed to re-vote and did not pass the criterion. For validity, the subgroup members expressed concern with the four years of "pooled" data and the multiple rounds of documentation reviews. Multiple reviews would not be considered a "gold standard" patient/encounter validity testing approach; therefore, that testing would be considered as patient/encounter reliability testing rather than validity testing. SMP members found the rounding review process confusing and potentially insufficient. The lead discussant also discussed data gaps that significantly affect the risk model, including the previously discussed osteoarthritis, no analysis of present on arrival (POA) diagnoses (e.g., pneumonia), and the treatment of approximately 20 percent of patients with post-surgical complications at locations that are different than where the surgery was performed. The subgroup also agreed to re-vote on the validity criterion but did not reach consensus. This measure is not eligible for the Standing Committee's revote during the fall 2021 evaluation cycle because appropriate levels of testing were not provided or otherwise did not meet NQF's minimum evaluation requirements. This new eCQM may be resubmitted in a future cycle for the SMP's reconsideration at the developer's discretion.

### #3650e Risk-Standardized Inpatient Respiratory Depression (IRD) Rate Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) eCQM (BWH)

#### *Measure Steward/Developer Representatives at the Meeting*

Patricia Dykes, Mica Bowen, Stuart Lipsitz

#### *Scientific Methods Panel Votes*

- **Reliability**: H-0; M-7; L-4; I-0 (Pass)
- **Validity**: H-0; M-3; L-8; I-2 (No Pass)

This new eCQM was pulled for SMP discussion because it did not pass on the validity criterion. In the preliminary analysis, the reliability test sample consisted of 17 total orthopedic groups: six from MGB and 11 from Cerner sites that perform between 25 and 1200 THA/TKA surgeries per year. Reliability testing was conducted at the patient/encounter level through the review of a small sample of 30 random patients to evaluate the accuracy of eCQM abstraction. The developer reported that all data elements abstracted by the eCQM matched with the information within the EHR. The developer also compared the sociodemographic characteristics of patients included in the test to validation samples and found no differences between sites or clinician groups. Some SMP subgroup members questioned whether reliability testing of sociodemographic characteristics was sufficient to demonstrate reliability. For accountable entity reliability testing, developers used a test-retest approach to examine the reliability of the predicted/expected ratios at the clinician group level with the same testing population by conducting a Spearman rank correlation of 0.767 between the two samples. The developer also estimated a low ICC between clinician groups and the ICC value at 0.069. SMP members expressed reliability concerns with the moderate correlation statistic compared to the very low ICC presented. They also expressed concerns with data collection over a four-year period for a measure specified with a single performance year, as well as concerns for potential overestimates of the Spearman correlation coefficient that used data across two calendar years instead of one. Reliability was not discussed because the subgroup agreed that the discussion focus should center on significant deficits for the validity testing results that did not pass the validity criterion.

The developer conducted patient/encounter level validity testing, citing the frequency of data elements needed for risk adjustment and data element agreement between manual chart review and EHR calculation. During the discussion, the SMP members stated that patient/encounter level validity testing of the gold standard was not conducted. Therefore, the conducted patient/encounter testing was an assessment of the *reliability* of the data elements. The developer also identified the presence of data inconsistencies and errors in both the EHR and manually abstracted data elements. Without patient/encounter validity testing, the SMP's validity assessment would depend on the results of the accountable-entity testing, for which the developers convened a Technical Expert Panel (TEP) to assess the face validity. The developer reported that three-sevenths (42.86 percent) of the TEP members agreed that the measure was actionable to improve quality of care. The SMP members noted the low results of conducted face validity testing and recognized that the developer's additional information clarified that the TEP wanted the measure's level of analysis to be at the facility level rather than the clinical group level. The lead discussant stated that there was no reason to refute the TEP's findings.

The developer risk-adjusted both the predicted and expected numerator events for age, gender, type of surgery (THA/TKA), insurance, race, household income, English as [a/the] primary language, smoking status, body mass index (BMI), and comorbidities. Several SMP members raised concerns with the conceptual rationale for the risk adjustment strategy, noting that the use of social risk factors (i.e., race, income, and insurance status) should not be used in this measure without a strong conceptual framework for why these might influence inpatient respiratory depression (IRD). Following the discussion, the SMP subgroup agreed to accept the low vote from their preliminary analyses. This measure is ineligible for the Standing Committee's revote during the fall 2021 evaluation cycle because an inappropriate methodology or testing approach was applied to demonstrate reliability and validity, and appropriate levels of testing were not provided or otherwise did not meet NQF's minimum evaluation requirements. This new eCQM may be resubmitted in a future cycle for the SMP's reconsideration at the developer's discretion.

### #3652e Risk-Standardized Prolonged Opioid Prescribing Rate Following Elective Primary Total Hip Arthroplasty (THA) and/or Total Knee Arthroplasty (TKA) eCQM (BWH)

#### *Measure Steward/Developer Representatives at the Meeting*

Patricia Dykes, Mica Bowen, Stuart Lipsitz

#### *Scientific Methods Panel Votes*

- **Reliability**: H-0; M-7; L-3; I-1 (Pass)
- **Validity**: H-2; M-5; L-4; I-0 (Pass)

This new eCQM measure was pulled for SMP discussion based on an SMP subgroup member's request regarding concerns with the conducted reliability and validity testing. The SMP subgroup discussed both reliability and validity. A vote for each criterion was rescheduled from the first to the second day of the

SMP web meeting after the subgroup members were guided to review potential unadjusted validity results in the submission. Prior to the second day of the SMP meeting, the measure developer withdrew the measure from further consideration during this evaluation cycle.

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

### *Measure Steward/Developer Representatives at the Meeting*

Elaine Breck, Ronen Rozenblum, Stephanie Singleton

### *Scientific Methods Panel Votes*

- **Reliability**: H-0; M-1; L-5; I-3 (No Pass)
- **Validity**: H-0; M-3; L-3; I-3 (No Pass)

This new measure was pulled for SMP discussion because it did not pass the reliability and validity criteria. In their preliminary analyses, the SMP subgroup members expressed significant concerns with the small sample sizes used for both reliability and validity testing at the accountable-entity level; three THA and four TKA clinician groups were used in the final testing. For a PRO-PM, reliability and validity must be demonstrated for both the instrument (i.e., at the patient/encounter level) and the performance measure score (i.e., at the accountable-entity level). The developer conducted patient/encounter level reliability testing with data from the Electronic Data Warehouse (EDW) and through manual chart review (n=68; 34 THA and 34 TKA patients). The developer stated that both data sources were prone to inaccuracies; therefore, they were not considered gold standard reviews. The alignment between the manual reviewers and the EDW were overall very strong with 97.1 percent agreement and a kappa value of 0.93. The developer stated that these results are above the 0.70 threshold, indicating strong agreement between the EDW data elements and manual abstraction data for the specified measure population, timeframe, and setting.  Reliability for the instrument data elements was not conducted. Instead, the developer cited published test-retest and internal consistency evaluations to assess reliability of both PRO-PM instruments or PROMs (i.e., hip disability and osteoarthritis outcome score [HOOS, JR] and knee injury and osteoarthritis outcome score [KOOS, JR]). Regarding the patient/encounter level reliability testing, SMP members also raised concerns that testing of the psychometric properties of the survey in practice as specified in the measure were not done; only testing of the denominator exclusions was conducted. Additionally, the numerator of the PRO-PM was not examined.

For both levels of reliability testing, developers conducted both patient/encounter and accountability entity level testing with sample sizes that were small and therefore contributed to a small signal-to-noise ratio (SNR). For validity testing, the small sample sizes contributed to an inability to draw conclusions about testing for known groups and weak results for discriminant validity testing. Prior to the meeting, the measure developer submitted a response acknowledging concerns about small sample sizes and low variability. The developer stated that delaying PRO-PM implementation would inhibit further testing and potentially widen already identified THA and TKA care gaps in assessing care goal achievement. The subgroup members also stated that the risk adjustment model showed inconsistent results for hip and knee and within the three levels for some of the models. For patient/encounter level validity testing, the developer did not present a gold standard for inter-rater reliability assessment and stated that this is because both the EDW and the manual chart review have the potential for inaccuracies. Without a gold standard, the subgroup members agreed that patient/encounter validity testing was not conducted. that patient/encounter validity testing was not conducted.

During the meeting, SMP members reiterated concerns with small sample sizes, and a lack of variability in testing results warranted a “not passed” rating. One SMP member expressed concern with the lack of correlational studies with the measure and patient quality of life post-surgery.  Another SMP member raised additional concerns with the developer’s missing data analysis, specifically that the submission did not address a steep drop in respondents in the pre- to post-surveys of more than 50 percent for both THA and TKA. SMP members asked the developer for available reasoning or conducted testing on the population of nonresponders, as this measure calculates the difference between pre- and post-surgery scores, and respondent data are needed to calculate measure results. The developer stated they did not conduct respondent/nonrespondent analyses but surmised the that the coronavirus disease 2019 (COVID-19) pandemic had a significant impact on the post-surgery response rate.

The SMP members agreed that no conclusions could be reached about the reliability of the measure due to small sample sizes and that the developer did not conduct sufficient testing of all data elements. The SMP recommended more testing in more clinician groups to contribute to a better understanding of how this operates at the entity level and testing for essential data elements at the patient/encounter level, especially the numerator. The SMP subgroup members agreed to uphold both of their preliminary votes to not pass the measure, given their lingering concerns. This measure is ineligible for the Standing Committee’s revote during the fall 2021 evaluation cycle because a methodology or testing approach was applied to demonstrate reliability and validity, and appropriate levels of testing were not provided or otherwise did not meet NQF’s minimum evaluation requirements. This new measure may be resubmitted in a future cycle for the SMP’s reconsideration at the developer’s discretion.

### #3639 Clinician-Level and Clinician Group-Level Total Hip Arthroplasty and/or Total Knee Arthroplasty (THA and TKA) Patient-Reported Outcome-Based Performance Measure (PRO-PM) (Centers for Medicare & Medicaid Services [CMS]/Yale Center for Outcomes Research & Evaluation [CORE])

#### *Measure Steward/Developer Representatives at the Meeting*

Doris Peter, Smitha Vellanky, Lisa Suter

#### *Scientific Methods Panel Votes*

- **Reliability**: H-3; M-3; L-1; I-2 (Pass)
- **Validity**: H-0; M-7; L-1; I-1 (Pass)

This new measure was pulled for SMP discussion as a request from one subgroup member who expressed concerns that the validity testing of the patient-reported outcome measure (PROM) used to collect the numerator data for the PRO-PM applied the reliability of comparator PROMs to PROMs in the measure. In their preliminary analyses, the SMP subgroup passed the measure on both reliability and validity, although they expressed concern with the validity testing as stated. As stated above for a PRO-PM, reliability and validity should be demonstrated for both the instrument (i.e., at the patient/encounter level) and the performance measure score (i.e., at the accountable-entity level). In the preliminary analyses, the subgroup members noted the developer used test-retest and internal consistency results from comparator PROMs (i.e., HOOS, JR and KOOS, JR) to determine reliability of the THA and TKA PROMs in this PRO-PM. The mean reliability score was 0.69 (SD 0.16) for clinicians with at least five cases. One SMP member raised concern with the response variation of social risk (i.e., race) and the potential underrepresentation of the experiences among certain racial groups in the sample.

For patient/encounter level validity testing, the developer evaluated both instruments using standardized response means and then compared against two other previously validated PROMs. During the SMP’s discussion, subgroup members noted concern with the ceiling effect in the patient/encounter level relating to the potential for measurable improvement of an entire standard deviation, which is inconsistent with the cited evidence. Concerns referenced the large 22-point PROM improvement threshold, which results in a substantial percentage of patients not meeting the target thresholds. One subgroup member stated that a high percentage of preoperative patients in the sample will automatically “fail” the measure based on very low or very high PROM scores. Another SMP member stated that the measure does not adequately account for patient comorbidities and their corresponding illness severities that might contribute to pre- and postoperative PROM responses. The developer explained that removing patients with high preoperative PROM scores reduces potentially unnecessary THA or TKA surgeries that could be managed medically. Additionally, the developer explained that from the orthopedics perspective, the ceiling effect is not concerning because it encourages clinicians and clinician groups to only offer surgery to patients with substantive symptom benefit potential.

For accountable -validity testing, developers conducted face validity by asking a 17-member TEP to respond to two statements related to the specified measure using a six-point scale. The SMP noted that of the two questions asked of the TEP, only seven of TEP members strongly agreed that “the PRO-PM as specified will provide valid assessment of improvement of functional status and pain following surgery” and only three of 17 strongly agreed that “the measure can be used to distinguish between better and worse quality of care among clinicians and clinician groups An SMP member stated that the second question, which addressed the entity level, had poorer results than the first, noting that only three TEP members strongly agreed. The SMP asked the developer to clarify why two individuals from the TEP disagreed with the questions asked. The developer stated their findings represent the 14 TEP members who strongly agreed or moderately agreed that the measure was valid. The developer stated the two dissenting TEP members had concerns with the method for incentivizing performance and wanted to see the measure used in a broader or different data sample.

After discussion, the SMP subgroup members agreed not to revote on this measure as it was determined that the validity issues were not concerning enough to change their preliminary votes which passed the measure. Therefore, upholding their previous votes, the SMP found the measure to be reliable and valid. The Surgery Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3667 Days at Home for Patients with Complex, Chronic Conditions (CMS/Yale CORE)

#### *Measure Steward/Developer Representatives at the Meeting*

Susannah Bernheim, Kyle Bagshaw, Jeph Herrin, Kyaw Sint Joe

#### *Scientific Methods Panel Votes*

- **Reliability**: H-5; M-6; L-0; I-0 (Pass)
- **Validity**: H-0; M-4; L-5; I-1 (Consensus Not Reached)

This new measure was pulled for discussion based on an SMP member’s concern with the validity testing approach, and the complexity and approach of the risk model. In their preliminary analyses, the subgroup members found the specifications confusing and occasionally arbitrary. Members reported the potential misalignment of concept presentations within the submission and noted the denominator statement lacked an explanation of the target population, conditions, settings, and other pertinent measure constructs information. They were also concerned that several concepts included in the submission were not documented as exclusions in the specifications, which both threatens the measure’s validity and may incentivize under-treatment of conditions potentially outside the locus of control of the accountable entity. The SMP also questioned whether the consideration of exclusions included (i.e., patients treated in emergency departments, admitted to acute care settings, and days after a death occurs), were always indicate low-quality care. Another SMP member expressed concerns with adjusting for transitions to the nursing home, which purports that moving from home to a nursing home, is always negative. Other concerning date elements included permanent nursing home admissions requiring skilled nursing care, which may include personal and community resources that are not be modifiable by the accountable entity. SMP members also noted that the unit of analysis reported in the measure vacillated between accountable care organizations (ACOs) and provider group.

For reliability testing, the developer conducted accountable entity-level testing using a split-half methodology with data from 2017-2018. They reported an ICC of 0.8326 for the final Days at Home outcome metric between the two samples. Beyond a "split-half" analysis, the form of ICC is not described. SMP members noted that the use of two years of data might not be appropriate, because patient assignment and ACO rules are modified annually. For validity, developers conducted accountable entity-level testing using construct validity with Pearson correlations to six other ACO-level measures hypothesizing that quality conceptually relates to excess days in care (EDIC) for patients with complex chronic diseases.  Pearson's correlations ranged between -0.549 and +0.048 resulting in a high inverse correlation for unplanned admissions (expected), moderate correlation with other measures, no correlation with fall risk, and an unexpected inverse correlation with patient experience. The developer reported poor correlations may result from testing against measures using smaller sample sizes and which were not risk adjusted for clinical variables. For accountable entity validity testing, the developer performed face validity testing of the computed measure score. A TEP consisted of 19 of 21 responding members who assessed whether the "The Days at Home measure, as specified, can be used to distinguish between better or worse performance at ACOs or provider groups." Two members indicated "strongly agree," 15 indicated "agree," and two indicated "somewhat agree." For the risk model, adjusted days at home calculates "excess days in care" using three risk models: 1) risk-adjusted days in acute care settings or SNFs among days alive in the year, 2) risk of mortality, and 3) risk of transition to nursing home. The EDIC updates are based on risk of death and risk of transition to nursing home care, and then averaged across each provider group to produce the final measure scores. Some SMP members noted that there are three different risk adjustment models used and expressed concerns about lack of clarity about whether/how they were combined to get a single score and the validity of the approach. One SMP member asked why primary death data was not used, instead of the presented death risk model. Further, SMP members had concerns with the model construction, which they agreed lacked vital adjustment and consideration for many variables without theoretical or empirical justifications and used arbitrary measure weighting. The c-statistic was 0.738 for the mortality model and 0.760 for nursing home transition. Deviance from R-squared was 0.170 for the EDIC model. Spearman rank correlation was 0.346 for more days in care. The subgroup members indicated the discrimination and calibration are generally acceptable but had concerns about the highest days in care decile, which raised further concerns related to outliers, (e.g., except for highest days in care decile). Members also expressed concerns with the results from the excess days and mortality and the method of combining nursing home transitions. The SMP members questioned the presence of meaningful differences in performance. The developer stated that a three-day difference is significant from a cost perspective, but one member noted that a difference in three days could reflect variables not included in the risk adjustment model or in residual effects not fully adjusted.  SMP members observed that it is not clear whether this equates to meaningful differences in quality of care. They questioned whether the measure would identify meaningful differences manifested for example, in differences in patient function or health-related quality of life. One SMP member was concerned that the developer did not present testing on between versus within ACO variance, adjusted for risk factors.

During the SMP meeting, the subgroup members focused their discussion on the validity testing. The lead discussant noted the developer submitted both face validity of the computed measure score and empirical accountable entity level testing that demonstrated correlations with endogenous measures or in a counterintuitive direction. The lead discussant questioned the process-outcome pathway that resulted in increased, rather than decreased, days in care, and the lack of exclusions for long-term nursing home residents prior to a measurement period, who have no chance of “at home” days defined in the specifications. SMP members discussed possible alternative ways in which the measure might have been constructed. The greatest concern was the development approach for days at home, and the mortality and nursing models. Faulty formulas in the approach may include doubling the EDIC estimates for enrolled ACOs and negative impacts to the penalty schematic. The developer stated the measure attempts to balance days at home with other unintended consequences. One member stated that measure may incentivize higher mortality with lower EDIC proportions. Other SMP members were concerned with the arbitrary and unexplained selection of weighting mortality days at 1.25 percent and the annual nursing home start date of January 1 that are not conceptually and empirically demonstrated or justified. The developers acknowledge these were not empirically assessed, but rather are subjective and selected by TEP recommendation. Another member questioned why nursing home stays are always perceived as negative? The developer confirmed this was TEP-recommended. Another member questioned if the measure should have been presented as a composite due to the multiple performance rates and the single overall score. NQF staff guided the subgroup to assess the measure as specified. The SMP found the measure to be reliable and revoted on the validity criterion but did not reach consensus. The Surgery Standing Committee will evaluate this measure in the fall 2021 cycle.

## Subgroup 2

During the meeting, subgroup 2 discussed one measure (#0689) and revoted on validity for measure and accepted the preliminary analysis vote for reliability without further discussion. The five remaining subgroup 2 measures (#3633e, #3662e, #3663e, #3665, and #3666) passed both the reliability and validity criteria, therefore, were not discussed as the SMP subgroup. The final results for the six measures evaluated by subgroup 2 are presented below.

### 0689 Percent of Residents Who Lose Too Much Weight (Long-Stay) (CMS/Acumen, LLC)

#### *Measure Steward/Developer Representatives at the Meeting*

Sriniketh Nagavarapu, Chang Lin

#### *Scientific Methods Panel Votes*

- **Reliability**: H-3; M-5; L-3; I-0 (Pass)
- **Validity**: H-0; M-4; L-5; I-0 (Consensus Not Reached)

This maintenance measure was pulled for SMP discussion as consensus was not reached on the validity criterion. In their preliminary analysis, the SMP subgroup noted the developers conducted reliability testing of the accountable entity score in both split-half reliability and beta-binomial signal-to-noise (SNR) testing. For split-half reliability testing, developers used the Spearman Rank Correlation and Pearson Correlation ($r = 0.64$, $\rho = 0.65$, ICC = 0.64, $p < .01$), suggesting moderate internal reliability. For the SNR, the relationship was also moderate (SNR = 0.76). Subgroup members sought to evaluate distributions of SNR reliability scores across facilities, which were not presented by developers. Several reviewers noted that less than half the facilities (44.6 percent) had no between quarter performance change, 20 percent had a greater than a three-decile between quarter performance change, and more than one third (37.4 percent) had a mean performance score where the CI did not include the national mean. During the meeting, the SMP subgroup members accepted the moderate rating result from the preliminary analysis and did not discuss that criterion.

Developers conducted both patient/encounter and accountable entity level validity testing. For accountable entity validity testing, the developers conducted multiple tests at the facility level including the following: 1) Convergent validity of publicly reported quality measures hypothesizing the facility's percentile rank would be "somewhat" consistent among the quality measures, correlations with other measures of nursing facility quality (i.e., three MDS quality measures and Facility Five-Star Rating, health inspections rating, and staffing levels (overall and for registered nurses (RNs)); 2) Variations by measure scores by states; 3) Seasonal variation based on non-facility related differences; 4) Stability analysis demonstrating relative facility decile ranking over time; and 5) Confidence interval analysis detailing performance significantly different from the national facility-level mean stratified by facility size. The results demonstrated: 1) For convergent validity, Spearman Correlation results in very weak correlations with the three quality measures: Percent of Residents Whose Ability to Move Independently Worsened (r = .113), Percent of Residents Whose Need for Help with Activities of Daily Living Has Increased (r = .108), and Percent of Residents Who Have Depressive Symptoms (r = .063). Correlations were negative between the facility-level weight loss measure score and the overall quality rating (r = -.143), health inspection rating (r = -.056), overall staffing level (r = -.029), and RN staffing (r = -.011); 2) State variation demonstrated a 4.8 percent overall variance with an average inter-quartile range of state-level scores of 3.8 percentage points; 3) Seasonal variation showed the highest weight loss in quarter one with progressively lower rates in quarters two and four; 4) Stability analysis based on two reporting periods by four rolling quarters showed 39.2 percent of facilities' percentile rankings were constant within the same decile, 37.9 percent changed rank within one decile, 15.4 percent changed rank within two deciles, and 7.5 percent changed rank by three or more deciles; 5) Confidence interval analysis demonstrated 28.7 percent of facilities had scores that were statistically significantly different from the national mean with 95 percent confidence, 15.4 percent were statistically significantly lower and 13.2 percent were statistically significantly higher than the national mean. Subgroup reviewers expressed concerns with convergent validity correlation results, citing very weak and negative correlations between the facility-level weight loss measure score and publicly reported measures and overall quality rating. Another subgroup member stated the low results may be common, although lower than usually seen, noting that overall nursing home quality and staffing have little impact on residents' likelihood of losing weight, while the resident's diagnoses and severity of illness may drive much of the weight loss. Multiple reviewers were concerned with the decision not to risk adjust the measure, specifically related to a resident's health status. The literature indicates there are potentially addressable risk factors for unintentional weight loss in long term care facility residents, including MDS score items for depression, Alzheimer's, cancer, Parkinson's disease, cognitive impairment, cardiac disorders, benign gastro diseases, and eating dependencies (i.e., chewing and swallowing issues) leading to 25 percent or more of food uneaten, and swallowing/chewing problems. The developers reported their attempts to develop a risk-adjusted model were unsuccessful and resulted in a low R-Squared value. Data on the covariate testing for the considered risk model were not provided by the developers ahead of the meeting, beyond age and White/Non-White race variables.

During the measure discussion, the lead discussant questioned whether the validity results met methodological standards and thresholds and stated a lack of confidence in understanding the conveyed message in the attempted validity testing and results. Specifically, hypothesized convergent validity correlations were very low and not convincing, state variations did not detail discernable insight, stability analysis demonstrated significant decile jumps between deciles (> 22.9 percent), seasonality did not demonstrate hypothesized findings, the confidence intervals above and below national means demonstrated variation. The information provided did not detail actionable quality differences between high- and low-performing users. Many reviewers discussed the lack of testing to support differences in facility populations at higher risk of unintended weight loss, specifically why MDS clinical data variables were not used in the testing or shown with risk adjustment considerations. The developers stated the measure will be risk-adjusted for these variables within an implemented payment model instead.

Subgroup members questioned how to vote on the criterion with varying results from the validity testing. The 2015 patient/encounter (i.e., data element) validity testing demonstrated high results and the subgroup 2021 accountable entity level validity testing did not adequately demonstrate validity. NQF staff guided the subgroup in a review of Algorithm 3. Guidance for Evaluating Validity in the NQF measure evaluation criteria with the highest possible criterion vote of moderate. For voting, subgroup members were guided to consider all the information presented by the developer in the current submission, the content of the preliminary analyses, and the measure discussion. The developer also questioned if the subgroup evaluated the current testing results during their preliminary analyses, or previous testing results, as some reviewers commented on “old testing” from 2015 in their preliminary analyses. One subgroup member confirmed the previous testing was reviewed, although after reviewing the new information, this member confirmed their validity result would not change. On day one, the validity criterion did not pass. To confirm the measure discussion focused on the applicable validity testing, NQF staff reviewed the measure discussion on day two of the meeting and requested that the subgroup members re-review the validity testing data from the current submission in preparation for day two of the web meeting. Another robust discussion occurred on day two of the web meeting. A revote for validity was taken and the measure did not reach consensus. The Patient Safety Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3633e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Level) (Alara Imaging/University of California San Francisco (UCSF))

This is electronic clinical quality measure (eCQM)

#### *Scientific Methods Panel Votes*

- **Reliability**: H-9; M-2; L-0; I-0 (Pass)
- **Validity**: H-5; M-6; L-0; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Patient Safety Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3662e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Clinician Group Level) (Alara Imaging/UCSF)

This is electronic clinical quality measure (eCQM)

#### *Scientific Methods Panel Votes*

- **Reliability**: H-8; M-3; L-0; I-0 (Pass)
- **Validity**: H-7; M-4; L-0; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Patient Safety Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3663e Excessive Radiation Dose or Inadequate Image Quality for Diagnostic Computed Tomography (CT) in Adults (Facility Level) (Alara Imaging/UCSF)

This is electronic clinical quality measure (eCQM)

#### *Scientific Methods Panel Votes*

- **Reliability**: H-9; M-2; L-0; I-0 (Pass)

- **Validity**: H-6; M-5; L-0; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Patient Safety Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3665 Ambulatory Palliative Care Patients' Experience of Feeling Heard and Understood (American Academy of Hospice and Palliative Medicine (AAHPM))

#### *Scientific Methods Panel Votes*

- **Reliability**: H-3; M-6; L-1; I-1 (Pass)
- **Validity**: H-3; M-5; L-3; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Geriatrics and Palliative Care Standing Committee will evaluate this measure in the fall 2021 cycle.

### 3666 Ambulatory Palliative Care Patients' Experience of Receiving Desired Help for Pain (AAHPM)

#### *Scientific Methods Panel Votes*

- **Reliability**: H-4; M-5; L-2; I-0 (Pass)
- **Validity**: H-2; M-6; L-3; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Geriatrics and Palliative Care Standing Committee will evaluate this measure in the fall 2021 cycle.

## Discussion of Overarching Methodological Issues Identified During Measure Evaluation

Throughout the two-day web meeting, SMP members identified and discussed numerous methodological and evaluation overarching themes for potential consideration at future advisory meetings. During this SMP measure evaluation meeting, three overarching issues were identified for possible deliberation. A brief summary of the issues is provided here.

#### *Scientific Acceptability Ratings*

During multiple measure discussions, SMP members sought direction from NQF staff related to the ratings for scientific acceptability criterion. Members stated the information provided in the narrative guidance and reliability and validity algorithms within NQF's measure evaluation criteria may differ from their expert opinions, experiences with measure evaluations, and testing expertise. Members requested to discuss potential updates and clarifications of the scientific acceptability algorithms in future advisory meetings. Specific areas of interest included whether to follow the algorithm exclusively when rating a criterion, or whether they have the latitude to consider the algorithms with other criterion guidance, the submission content, meeting discussions, and their expertise. Another area of requested guidance is clarification on divergent testing results (e.g., high patient/ encounter results with low accountable entity results for reliability results). During the meeting's measure reviews, members were guided to consider all presented and discussed information to the specified measures within criterion evaluations.

#### *Face Validity Guidance*

SMP members expressed concern with sufficiency of submitted face validity that is limited to an assessment of "computed measures scores" by asking the TEP one or two questions. Some members stated that developers may require additional technical assistance and testing education related to face validity, such as how to construct adequate tests of face validity, identify process measures as validity comparators or correlates when correlate outcome measures are difficult to identify, and the use of predictive validity tests to assess whether the outcome measure is appropriate to assess performance. During multiple measure reviews, the members questioned whether accepting face validity alone for accountable entity-level validity testing is appropriate, especially for measure used for high-stakes uses, such as CMS measure programs and value-based payment (VBP). NQF staff clarified that face validity is currently an acceptable form to demonstrate computed measure score or accountable entity level validity testing for newly submitted measures. They also clarified that a level of agreement decision is a criterion that the SMP may or may not accept. The SMP asked to address these at future advisory meetings.

### *Small sample sizes*

Within multiple measure discussions, SMP members asked for guidance on how to assess measures with small sample sizes as NQF is not prescriptive in defining minimum requirement for reliability, validity testing, and risk modeling. SMP members noted the significant challenges in evaluating measures with small testing volumes, as well as developer concerns with the difficulty of capturing captures data that inhibits adequate scientific acceptability assessment. They also stated inherent challenges in considering the generalizability of measure within and between applicable levels of analysis, especially for measures specified at the facility and larger levels. During the measure discussions, members were guided to use the available information at hand to rate the measures.

### Public Comment

No public or NQF member comments were provided during the measure evaluation meeting.

## Next Steps

Ms. Ingber reviewed the SMP next steps and reminders for the measures reviewed by SMP members during this cycle. NQF staff will inform developers and Standing Committees of the SMP's discussion and votes. Measures that passed both reliability and validity or for which consensus was not reached will be considered by the relevant Standing Committees in the fall 2021 evaluation cycle. According to NQF endorsement guidance, eligible measures that did not pass the SMP's vote may be pulled for discussion and revote by the relevant Standing Committee. However, a measure is *not* eligible for Standing Committee discussion and revote if any of the following are true: (1) Inappropriate methodology or testing approach was applied to demonstrate reliability or validity, (2) Incorrect calculations or formulas were used for testing, (3) Description of testing approach, results, or data is insufficient for the SMP to apply the criteria, and (4) Appropriate levels of testing were not provided or otherwise did not meet NQF's minimum evaluation requirements. As discussed in individual preceding measure evaluations, measures #3638 and #3649e, and #3650e are not eligible for Standing Committee revote. Measures moving forward in the fall 2021 evaluation cycle will be reviewed by their respective Standing Committees in February 2022 and discussed by the Consensus Standards Approval Committee (CSAC) in July 2022.

The SMP will convene via web meeting on December 14, 2021, to continue a discussion on acceptable reliability thresholds and CSAC updates that were not discussed during the meeting due to time constraints. Other methodological evaluation topics identified by the SMP during this and previous web meetings may also be discussed during the next SMP advisory web meeting.