NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel Measure Evaluation Web Meeting – Spring 2021

The National Quality Forum (NQF) convened the Scientific Methods Panel (SMP) on March 30-31, 2021, for a discussion of the scientific acceptability (i.e., reliability and validity) of several complex measures submitted to the spring 2021 evaluation cycle. Of the 29 measures reviewed by the SMP during this cycle, 12 measures were discussed at the meeting, including those for which the review subgroup (1) did not reach consensus and/or had major areas of concern in their preliminary evaluations, (2) did not initially pass the SMP's evaluation but for which measure developers provided additional information, and (3) were otherwise pulled for discussion by the SMP subgroup[1] members or NQF staff. This meeting summary includes brief summaries of the 12 measures, the overarching methodological issues discussed during the web meeting, and the voting results for the 15 measures not discussed during the web meeting. Two measures were withdrawn from the cycle prior to the meeting.

Welcome, Introductions, and Review of Meeting Objectives

Sai Ma, NQF managing director and senior technical expert, welcomed the members of the SMP, measure developers, other NQF staff, and members of the public to the web meeting. Chris Queram, NQF interim president and CEO, Sheri Winsper, NQF senior vice president of Quality Measurement, and SMP Co-Chairs David Nerenz and Christie Teigland also provided opening remarks. Hannah Ingber, NQF senior analyst, asked SMP members to introduce themselves and to 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 spring 2021 SMP member recusals document. In short, Dr. Nerenz identified a conflict of interest with measure #0500; SMP member Samuel Simon identified conflicts of interest with measures #0500, #2860, #2902, #2903, and #2904; SMP member Zhenqiu Lin identified conflicts of interest with measures #2431, #2436, #2579, #2880, #2881, #2882, and #3612; SMP member Patrick Romano identified a conflict of interest with measure #3501; SMP member Susan White identified a conflict of interest with measure #3188; SMP member Sean O'Brien identified a conflict of interest with measure #3610; and SMP member Matt Austin identified a conflict of interest with measure #3614. These SMP members did not participate verbally, via the web meeting chat, or through email during discussion of the respective measures. Dr. Ma then described the process for the measure discussions and highlighted key NQF evaluation criteria.

Measure Evaluation

Thirty-eight measures were submitted during the spring 2021 cycle, 29 of which are complex measures; they were reviewed by the SMP for their scientific acceptability. Measure evaluations during the meeting were based on the preliminary analyses performed by assigned members of the SMP. Each SMP member was assigned to one of three subgroups, and each subgroup was assigned nine or 10 of the 29

[1] Subgroups are determined by the NQF SMP team with input from the SMP co-chairs. NQF assigns measures to subgroup members for evaluation based on panelists' relevant expertise, interests, measure specific disclosures of interest, and Standing Committee membership.

measures being reviewed this cycle. Subgroup members then performed in-depth reviews and analyses of their assigned measures. Developers received these preliminary analyses prior to the web meeting and were given an opportunity to submit written responses to concerns expressed by SMP reviewers. These responses were provided to the SMP prior to the meeting to prepare for their discussion and subsequent voting by subgroups on the measures during the meeting. Two measures (#2967 and #2689) were withdrawn from this review cycle by the developers after the preliminary evaluation but prior to the web meeting.

During the meeting, the SMP evaluated the reliability and/or validity for 12 measures based on their preliminary analyses and additional information submitted for consideration by the developers. For each measure discussed, NQF staff described the measure, noted the preliminary evaluation ratings of the subgroup, and highlighted the criterion (or criteria) for which there was a lack of consensus and/or major areas of concern. To ensure reliability is consistently evaluated across the subgroups, any measure that has a reliability result at or below 0.50 was pulled for SMP discussion. Drs. Nerenz and Teigland, SMP co-chairs, facilitated the remainder of the discussion. Lead discussants from the subgroup first summarized the primary concerns of the subgroup. Other subgroup members then made additional comments. Following these concerns and comments, NQF staff invited measure developers to provide brief responses to the concerns raised by the subgroup members and to summarize their written response, if provided. Next, the co-chairs invited comments or additional questions from other SMP members, and developers were invited to respond. A quorum of greater than or equal to 66 percent (i.e., at least six SMP members) was achieved for all subgroup votes during the meeting. 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 by the respective SMP subgroups.

The remaining 15 measures evaluated by the SMP in the spring 2021 cycle were not discussed during the meeting because subgroup members reached consensus on the ratings, and the measures were not otherwise pulled for discussion. For these measures, the subgroup's preliminary analyses serve as the final SMP assessment of scientific acceptability for the 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 eight measures (#2880, #2881, #2882, #3188, #3612, #3615, #3616, and #3622). This subgroup re-voted on validity for measures #2880, #2882, #3188, and #3622, and on the reliability and validity of measures #2881 and #3612. To ensure reliability is consistently evaluated across the subgroups, the SMP and NQF decided to discuss any measure that has a reliability result lower than 0.50; therefore, measures #3615 and #3616 were pulled for SMP discussion. After further discussion, the subgroup accepted the preliminary analysis decisions for passing reliability for measures #3615 and #3616. The subgroup accepted the preliminary analysis decisions for measure #2860 without further discussion. One of the 10 measures evaluated by this subgroup was withdrawn from this review cycle prior to the meeting (#2967). The final results for the remaining nine measures evaluated by subgroup 1 are presented below.

#2880 Excess Days in Acute Care (EDAC) After Hospitalization for Heart Failure (HF) (Yale Center for Outcomes Research and Evaluation (CORE)/Centers for Medicare & Medicaid Services (CMS))

Measure Developer/Steward Representatives at the Meeting

Jackie Grady, Doris Peter, Duwa Amin, Lisa Sutter, Huihui Yu, James Poyer

Scientific Methods Panel Votes

- Reliability: H-0; M-8; L-1; I-0 (Pass)
- Validity: H-0; M-7; L-0; I-1 (Pass)

In their preliminary analyses, the SMP noted minor concerns with the reliability of the measure and passed it with a moderate rating. The SMP expressed two concerns with the validity of the measure. The first involved the choice of variables for the construct validity, noting the potential for endogeneity due to the overlapping readmission events between the EDAC measure and the Star Ratings measures, as the same readmission events are included in both measures. The second concern centered on the c-statistic of the risk adjustment model, noting that it was low (0.59).

The SMP acknowledged the developer's response to concerns regarding the overlapping readmission events. The developer provided updated testing results by removing the comparator measure from the Star Rating Readmission Group score before analyzing the correlation and by removing the entire Readmission Group score. The developer noted that a moderate correlation remains (r = -0.349 versus -0.399) in the expected direction between the EDAC measure and Star Ratings, even after removing the overlapping measure. The developer also found a moderate association (r = -0.457 versus -0.579), albeit weaker, between the EDAC measure and Star Ratings after removing the entire Readmission Group from Star Ratings.

Regarding the risk adjustment model, the SMP debated whether updates to the risk adjustment model should include creating both training and validation data sets, as some SMP members argued that this is standard in evaluating model performance for risk adjustment models. The SMP noted that this occurred during the initial development of the risk adjustment model for NQF #2880, but it was not done for the updates to the current model. An SMP member also argued that since the model now uses International Classification of Diseases (ICD)-10 codes instead of ICD-9 codes, it should be treated as a new model. This was the case for NQF #2881 and NQF #2882 as well. The developer confirmed that for ongoing model performance, they do not have a development (training) and validation set, as they are not reselecting risk variables every year. Instead, they simply recalculate the beta coefficients of the same risk variables. Some SMP members were not concerned by the lack of a validation data set for the updated model since the risk variables have not changed. The SMP agreed that there is no consensus on this issue in the evaluation guidance and that it should be discussed at a future SMP advisory meeting. Lastly, the SMP discussed why the risk adjustment model and claims data may not be good discriminators for admission based on the low c-statistic. The developer explained that various factors could help improve the model in which the developer does not have access, such as functional status. The developer added that when this measure was initially developed, it was modeled from data registries and that the clinical risk models based on claims do very poorly. The developer asked SMP members for guidance on how to proceed with optimal validation testing from their perspectives. An SMP member mentioned that including additional factors (e.g., lab results) may improve the model and that there is available evidence for employing machine learning (ML) that may also improve the model. The developer mentioned that they have conducted analyses with these measures using ML techniques and have seen model improvements with the mortality measures rather than the readmissions measures and that they will continue these analyses to learn more. The SMP re-voted after the discussion and passed the measure on validity. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the spring 2021 cycle.

#2881 Excess Days in Acute Care (EDAC) After Hospitalization for Acute Myocardial Infarction (AMI) (Yale CORE/CMS)

Measure Developer/Steward Representatives at the Meeting

Jackie Grady, Doris Peter, Duwa Amin, Lisa Sutter, Huihui Yu, James Poyer

Scientific Methods Panel Votes

- Reliability: H-0; M-3; L-5; I-0 (Not Pass)
- Validity: H-0; M-7; L-0; I-1 (Pass)

In their preliminary analyses, the SMP did not reach consensus on reliability and raised concerns with the split-sample intraclass correlation coefficient (ICC) values, which were generally low, specifically less than or equal to 0.40 for roughly three-fourths of the hospitals (i.e., those with ≤50 admissions), indicating low-to-moderate agreement between the split-samples. The ICC was 0.38 for hospitals with greater than or equal to 25 admissions and 0.63 for hospitals with greater than or equal to 300 admissions. For validity, the SMP raised similar concerns with NQF #2880 and NQF #2882, namely that the c-statistic was low (0.60) and that the variables chosen for construct validity were inappropriate due to the potential endogeneity.

Based on the results of the preliminary analyses, the SMP discussed the reliability and validity testing. During the meeting, the SMP considered the developer's response regarding the underlying cause of the lower reliability, such as smaller samples that are difficult to split into two groups with varied hospital and beneficiary characteristics of the target populations; as a result, they are not completely attributable to the measure. An SMP member commented that reliability results are less than the other two EDAC measures that the SMP passed on reliability (NQF #2880 and NQF #2882). The SMP member recommended that the minimum case volume of 25 should be re-evaluated to a higher cutoff number, such as 100 cases. The developer mentioned that increasing the volume size would lead to fewer hospitals being included in the measure and that these are measure and program implementation decisions. The measure steward, CMS, commented that they appreciate the recommendation for increasing the case volume due to the reliability concerns. However, there is a tradeoff between increasing the case volume and losing the number of eligible hospitals from the measure. The SMP re-voted after the discussion and did not pass the measure on reliability.

With respect to validity, the SMP did not have anything further to add, as the concerns were parallel to NQF #2880 and NQF #2882, which the developer previously addressed. The SMP re-voted and passed the measure on validity. The All-Cause Admissions and Readmissions Standing Committee can evaluate this measure in the spring 2021 cycle if a Standing Committee member pulls this measure for discussion.

#2882 Excess Days in Acute Care (EDAC) After Hospitalization for Pneumonia (Yale CORE/CMS)

Measure Developer/Steward Representatives at the Meeting

Jackie Grady, Doris Peter, Duwa Amin, Lisa Sutter, Huihui Yu, James Poyer

Scientific Methods Panel Votes

- Reliability: H-1; M-8; L-0; I-0 (Pass)
- Validity: H-0; M-7; L-0; I-1 (Pass)

In their preliminary analyses, the SMP noted minor concerns with the reliability of the measure and passed it with a moderate rating. Although the SMP recognized the method of validity testing for NQF #2880 was clear, there were some concerns with the c-statistic, noting that it was low (0.62). Additional concerns were raised regarding the choice of variables for the construct validity, noting the potential for

endogeneity due to the overlapping readmission events between the EDAC measure and the other measures, as the same readmission events are included in both measures.

Based on the results of the preliminary analyses, the SMP discussed the validity testing. During the meeting, the SMP mentioned that the concerns with NQF #2882 were parallel to NQF #2880. The SMP also acknowledged that in their response, the developer conducted additional testing to address the SMP's concerns with respect to overlapping readmission events and found similar results for NQF #2880. One SMP member asked what processes were used by developers to identify process measures for validity correlates to outcome measures. This same member noted that readmission rates could be helpful to discriminate hospital performance for this measure. The developer stated they would consider additional empirical testing and correlations to process measures for future maintenance cycles. The SMP agreed that clearer guidance on the considerations of process measure correlations for validation should be provided for future cycles and should be added to future SMP advisory meeting discussions. The SMP re-voted after the discussion and passed the measure on validity. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the spring 2021 cycle.

With respect to validity, the SMP stated they did not have anything further to add, and any concerns were parallel to NQF #2880 and NQF #2882, which the developer previously addressed. The SMP re-voted and passed the measure on validity. The All-Cause Admissions and Readmissions Standing Committee can evaluate this measure in the spring 2021 cycle if a Standing Committee member pulls this measure for discussion.

#3188 30-Day Unplanned Readmissions for Cancer Patients (Alliance of Dedicated Cancer Centers)

Measure Developer/Steward Representatives at the Meeting

Kristen McNiff, Karen Fields

Scientific Methods Panel Votes362

- Reliability: H-0; M-7; L-2; I-0 (Pass)
- Validity: H-0; M-2; L-3; I-2 (Not Pass)

In their preliminary analyses, the SMP reviewers questioned the use of a 50-admission cutoff in risk-adjusted split sample testing, although the measure specification has no sampling requirements. The SMP passed the measure on reliability with a moderate rating. The SMP did not pass the measure on validity, noting that two of the risk factors (i.e., Length of Stay [LOS] >3 days and Intensive Care Unit [ICU] stay) are not present at the start of care, and there is a well-studied pathway connecting shorter stays with higher readmission risk. The SMP further noted that although the overall c-statistic (c=0.711) is acceptable, the calibration in the highest risk group seems poor. Lastly, the SMP raised concerns with use of the correlation between this measure and the CMS Hospital-Wide All-Cause Readmission (HWR) measure (#1789) for score-level validity testing. Specific concerns included endogeneity and intrinsic correlation between these measures.

During the meeting, the developer discussed the ongoing gaps that persist in cancer quality measures. They noted that #1789 specifically excludes patients admitted for medical treatment of cancer (i.e., patients with a principal diagnosis of cancer unless undergoing a major surgical procedure) and that the SMP reviewers may not have been aware of the presence of these exclusions in HWR. They also stated that #1789 excludes index admissions to the nation's Prospective Payment System (PPS)-exempt cancer hospitals. Regarding the risk adjustment, the SMP mentioned that the model should be limited to factors that were present at the start of care (e.g., limited access to care and the inability of patients to access hospital data on performance) and should not include post-admission factors (e.g., LOS and ICU stays). The developer explained that for this maintenance submission, they used the same variables that

were used in the initial endorsement of this measure. An SMP member commented that the c-statistic is unusually high for readmission models, which may be due to risk adjustment for post-admission factors that are not usually included in the risk model (e.g., LOS and disposition to hospice) and that discharging patients quickly could lead to increased readmissions. The SMP re-voted after the discussion and did not pass this measure on validity. The All-Cause Admissions and Readmissions Standing Committee can evaluate this measure in the spring 2021 cycle if a Standing Committee member pulls it for discussion.

#3612 Risk-Standardized Acute Cardiovascular-Related Hospital Admission Rates for Patients With Heart Failure Under the Merit-Based Incentive Payment System (Yale CORE/CMS)

Measure Developer/Steward Representatives at the Meeting

Kasia Lipska

Scientific Methods Panel Votes

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

In their preliminary analyses, the SMP did not reach consensus on either reliability or validity. For reliability, the SMP raised concerns that the reliability tests are not conducted and presented for clinical groups and individual clinicians separately. They also noted the testing sample sizes were framed to meet reliability thresholds of 0.4 and 0.5 of 21 and 32 patients, respectively, with higher cutoff points generally increasing reliability results. However, only 24 percent of providers had 21 eligible cases. It was also unclear whether the provider-to-patient ratio (i.e., whether higher volume practices have a large number of clinicians who are each seeing a small number of patients versus individual providers with a high volume of patients) was included in the analysis. Regarding validity, the SMP generally agreed that the face validity was adequately established, but they questioned why *race* was not included in the risk adjustment analysis and why *dual eligibility* was not included in the final risk adjustment model, as they both can affect the outcome.

During the meeting, the developer commented that under the Merit-Based Incentive Payment System (MIPS), clinicians annually select whether to report as individuals, as part of a group, or as both. The group includes both solo clinicians (i.e., clinicians opting not to report with other clinicians under MIPS) and groups of clinicians who have chosen to report their quality under a common tax identification number (TIN). Therefore, testing results include both individual clinicians and clinician groups, consistent with how the MIPS program evaluates quality. Among TINs with a case volume of at least 21 HF patients (when reliability of 0.4 is reached), 31.8 percent were solo clinicians. An SMP member questioned whether this measure is only to be used at a minimum of 21 cases, as there were concerns that this would lead to overrepresentation of larger practices. The developer commented that the 21 minimum case volume was established to reach the reliability threshold of 0.4, although no sampling minimums are specified. The developer stated that the MIPS program will set the minimum case volume during rulemaking. The SMP re-voted after the discussion and passed the measure on reliability.

For validity, one SMP member raised some concerns with the potential response bias since not all Technical Expert Panel (TEP) members responded to the survey for face validity. Additionally, for both the TEP and the clinician committee that reviewed the measure, it is unclear what these groups suggested or recommended for the measure. In response, the developer explained that this was a multiyear TEP, and not all members were active throughout the time of development. Only those who remained on the TEP responded to the survey. With respect to the feedback, this measure underwent multiple revisions with input from both groups (TEP and clinician committee), such as excluding some high-risk HF patients due to clustering of patients to certain clinicians, conditions, or devices (e.g.,

pacemakers, end-stage renal disease [ESRD], and systolic HF), which all could lead to higher readmissions. The developer also stated that due to CMS's request, the measure is not risk-adjusted for race. The SMP re-voted after the discussion and passed the measure on validity. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the spring 2021 cycle.

#3615 Unsafe Opioid Prescriptions at the Prescriber Group Level (University of Michigan Kidney Epidemiology and Cost Center (UM-KECC))

Measure Developer/Steward Representatives at the Meeting

Jonathan Segal

Scientific Methods Panel Votes

- Reliability: H-6; M-1; L-1; I-1 (Pass)
- Validity: H-2; M-4; L-1; I-2 (Pass)

Both this measure and the subsequent measure (#3516) passed but were pulled for discussion by an SMP member, specifically to address an overarching question: To what extent is the validity analysis confounded by an unmeasured case mix, considering that dialysis physicians with sicker patients (e.g., those with comorbid cancer) have higher mortality rates, hospitalization rates, and opioid use? Therefore, the two measures were discussed concurrently.

In their preliminary analyses, the SMP noted that assessing the correlation between #3615 and hospitalization and mortality is an appropriate validity test, given the provided discussion of the literature. However, no specific correlation test is specified, and it appears that the relationships are stated with descriptive statistics.

During the meeting, the SMP discussed the use of a risk adjustment model for a process measure. The SMP noted that it would be more appropriate for risks to be made into exclusions (e.g., cancer). SMP members also noted that other endogenous factors (e.g., drug dependence, substance use disorder [SUD], anxiety disorders, and previous opioid poisoning) may increase risk and are confounders that may be difficult to understand or differentiate. The risk adjustment model was noted as appropriate in terms of performance statistics but lacked an underlying theory to justify the selection of factors for the model.

The SMP expressed concerns that the validation of the measure is based on dividing provider groups into tertiles that showed the top tertile with a failure rate of over 46 percent, the middle tertile between 30 and 36 percent, and the best tertile under 30 percent. The submission noted that patients in the worst performing tertile have a slightly higher hospitalization odds ratio (1.49 versus 1.41) and a few more hospital days per year (6.1 versus 4.1), as well a higher death rate. They also noted these findings were reported under an unadjusted analysis when the developer had suggested that risk adjustment is essential for the measure's application. The SMP elected not to re-vote on the measure but will pass along the concerns to the Renal Standing Committee.

#3616 Unsafe Opioid Prescriptions at the Dialysis Practitioner Group Level (UM-KECC)

Measure Developer/Steward Representatives at the Meeting

Jonathan Segal

Scientific Methods Panel Votes

- Reliability: H-1; M-6; L-1; I-1 (Pass)
- Validity: H-1; M-5; L-1; I-2 (Pass)

Both this measure and the previous measure (#3615) passed but were pulled for discussion by an SMP member, specifically to address an overarching question: To what extent is the validity analysis confounded by an unmeasured case mix, considering that dialysis physicians with sicker patients (e.g., those with comorbid cancer) have higher mortality rates, hospitalization rates, and opioid use?

In their preliminary analyses, the SMP noted that assessing the correlation between #3616 and hospitalization and mortality is an appropriate validity test, given the provided discussion of the literature. However, no specific correlation test is specified. The relationships are stated with descriptive statistics.

The SMP's discussion from the previous measure was carried over to this measure. The SMP noted the use of a risk adjustment model for a process measure, specifically that it would be more appropriate for risks to be made into exclusions (e.g., cancer). Furthermore, SMP members noted that other endogenous factors (e.g., drug dependence, SUD, anxiety disorders, and previous opioid poisoning) may increase risk and are confounders that may be difficult to understand or differentiate. The risk adjustment model was noted as appropriate in terms of performance statistics but lacked an underlying theory to justify the selection of factors for the model. The SMP elected not to re-vote on the measure but will pass along the concerns to the Renal Standing Committee.

#3622 National Core Indicators for Intellectual and Developmental Disabilities (ID/DD) Home and Community-Based Services (HCBS) Measures (Human Services Research Institute)

Measure Developer/Steward Representatives at the Meeting

Henan Li, Nilufer Isvan, Alixe Bonardi

Scientific Methods Panel Votes

- Reliability: H-3; M-3; L-2; I-1 (Pass)
- Validity: H-0; M-4; L-0; I-3 (Consensus Not Reached)

The measure passed on reliability but did not pass on validity during the review prior to the meeting. Therefore, the focus of the discussion was validity.

In their preliminary analyses, the SMP noted that the submission was not complete in the data element validity testing because the developer had only listed references to studies without appropriately summarizing their results; hence, there was no data element validity evaluation conducted by SMP reviewers. It was noted that none of the risk factors for this risk-adjusted measure were tested. Furthermore, the SMP noted that the developer's testing of performance score reliability at the state level was not optimal because all of the constructs are estimated based on the same survey, suggesting that any validity issues that affect the entire survey in a consistent manner are likely to lead to exaggerated correlations. The Committee suggested that analyses with external measures of quality in a comparable quality domain would have been more appropriate. The SMP further noted that the results of the Pearson product-moment correlation analyses are difficult to interpret because the theoretical relationship between the correlates chosen was not provided nor whether only statistically significant associations were returned/presented.

During the meeting, the SMP discussed a missing theory of quality between the 14 measures under the measure heading. The SMP also pointed out the possibility of a confounding influence of other factors. By way of example, they pointed to the first measure related to community job goals. The argument is that because urban settings provide greater job opportunities, one would expect a correlation at the state level between the percent of people who live in urban areas and the score of the measure, and indeed there is such a correlation. The Committee expressed uncertainty regarding whether this

establishes the validity of the measure or whether it may be a confounder that ought to be adjusted for because urbanicity is not a dimension of quality. The Committee asked the developer to describe how the pattern of relationships between each of the measures that the developer described establishes the validity of each of the measures. The Committee emphasized that the submission would benefit from a clear explanation of the quality construct for the measure. The developer noted that they had provided information related to this matter in the responses to the SMP's concerns, specifically responses that did the following: (1) provide theoretical/hypothetical context for the reported Pearson correlation coefficients, (2) correlate measures with external data, (3) report complete correlation results with proper corrections, and (4) provide information about # Person-Centered Planning (PCP)-1 (Community Job Goal), #PCP-3 (ADL Goal), # Community Inclusion (CI)-1 (Social Connectedness), and #CI-3 (Transportation Availability Scale). The developer articulated directional hypotheses for expected associations among measures and only tested those hypotheses, noting that all 14 measures were supported in at least one hypothesis. The developer provided a table that summarized the results within their response. The developer further noted that the home and community-based services (HCBS) report developed by NQF in 2016 provides a theoretical framework for HCBS quality and expressed that the measure itself is aligned with that quality framework. As an example, the developer suggested that high quality HCBS might provide a person-driven system to optimize individual choice, which aligns with the Choice and Control quality domain described in the 2016 HCBS report.

The SMP further noted that the calibration results for two of the measures in the case mix adjustment contained large discrepancies between observed and predicted values across deciles and predictive risk. The developer noted that a deeper understanding of this would require additional analysis, but the life decision scale, in particular, is made of several factors that are actually stand-alone instrument items; they also noted a tradeoff between more inclusion or better availability of the score. The SMP expressed concern about variability between interviewers across states. The developer noted that there are training regimens and train the trainer programs that aim to reduce variability between interviewers administering the survey.

#2860 Thirty-Day All-Cause Unplanned Readmission Following Psychiatric Hospitalization in an Inpatient Psychiatric Facility (IPF) (Yale CORE/CMS)

Scientific Methods Panel Votes

- Reliability: H-1, M-8, L-0, I-0 (Pass)
- Validity: H-1, M-6, L-1, I-1 (Pass)

Subgroup members found the measure to be both reliable and valid. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the spring 2021 cycle.

Subgroup 2

During the meeting, subgroup 2 discussed and re-voted on validity for one measure: #3614. The subgroup accepted the preliminary analysis decisions for nine measures (#1598, #1604, #2431, #2436, #2579, #3610, #3623, #3625, and #3626) without further discussion. The final results for the 10 measures evaluated by subgroup 2 are presented below.

#3614 Hospitalization After Release With Missed Dizzy Stroke (H.A.R.M Dizzy-Stroke) (Armstrong Institute for Patient Safety and Quality at Johns Hopkins University)

Measure Developer/Steward Representatives at the Meeting

Matt Austin, David Newman-Toker

Scientific Methods Panel Votes

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

In their preliminary analyses, the SMP noted minor concerns with the reliability of this measure and passed it with a moderate rating. For validity, the SMP expressed concerns regarding the data element validity for both the numerator and denominator, the risk adjustment methodology, and whether the measure would show meaningful differences between hospitals, given the need for a three-year measurement period and how providers would operationalize it in quality improvement efforts.

Regarding the validity of the data elements, the measure developer provided substantial validity testing for stroke, as well as coding validity for benign dizziness with both ICD-9 and ICD-10 codes. The SMP recognized the high rates for positive predictive value and negative predictive value from both validity testing approaches. The measure developer noted that coding for stroke and dizziness was consistent, and results were seen with multiple data sets and peer-reviewed publications. In addition, there was good face validity that posterior strokes (which present as dizziness) are more likely to be missed than anterior strokes. They also noted approximately five million emergency department (ED) visits occur for dizziness. Three to 5 percent of these patients will have a stroke, and some of these strokes are missed. Specifically, among patients with dizziness and stroke, about 40 percent are misdiagnosed. This measure helps to identify these misses to guide improvement.

Regarding risk adjustment, the SMP noted this measure is not risk-adjusted in a traditional sense. The statistical approach evaluates the observed, short-term stroke risk within 30 days minus the long-term expected risk. The long-term expected rate is estimated in the same patients using the 30-day rate of stroke admission during the period of 91-360 days post-discharge. The risk difference then quantifies the "excess" short-term stroke rate (attributable risk) due to misdiagnosis. This approach intrinsically captures hospital, patient, and social risk factors, given the patient as their own internal control. The measure developer noted a peak in stroke incidence within the first 30 days and then a linear phase. The linear phase reflects the stable, longer-term risk, which aligns with the risk for major stroke after minor stroke or transient ischemic attack.

The SMP also noted concerns related to the measure's ability to show a meaningful difference between hospitals since 65 percent of hospitals in the sample performed better than the national average, and no hospitals were worse than the national average due to the sample the developer used to develop and test this measure. The measure developer noted this was a limitation of the data set used (Medicare Fee-for-Service data from 5,000 hospitals and only 20 percent of all ED visits to the hospital). Additional data sets, such as an all-payer claims database or the Medicare 100 percent sample, would show better discrimination between high- and low-performing hospitals, given the larger sample sizes. They noted that rural hospitals are at greater risk of missing these events, but these hospitals also see fewer patients; therefore, they do not have enough data points to be included in the analysis. The measure developer noted this measure could be used in value-based purchasing in larger facilities where measure precision could improve and should be used in smaller facilities for quality improvement. The SMP re-voted after the discussion and passed the measure on validity with a moderate rating. The measure will move forward to be reviewed by the Neurology Standing Committee in the spring 2021 cycle.

#1598 Total Resource Use Population-Based PMPM Index (HealthPartners)

Scientific Methods Panel Votes

- Reliability: H-4; M-3; L-0; I-2 (Pass)

- Validity: H-4; M-2; L-1; I-2 (Pass)

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#1604 Total Cost of Care Population-Based PMPM Index (HealthPartners)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#2431 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Acute Myocardial Infarction (AMI) (Yale CORE/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#2436 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Heart Failure (HF) (Yale CORE/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#2579 Hospital-Level, Risk-Standardized Payment Associated With a 30-Day Episode-of-Care for Pneumonia (PN) (Yale CORE/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#3610 30-Day Risk-Standardized Morbidity and Mortality Composite Following Transcatheter Aortic Valve Replacement (TAVR) (American College of Cardiology)

Scientific Methods Panel Votes

- Reliability: H-0; M-7; L-1; I-0 (Pass)
- Validity: H-3; M-5; L-0; I-0 (Pass)
- Composite Construction: H-3; M-3; L-1; I-1 (Pass)

Subgroup members found the measure to be both reliable and valid. The Cardiovascular Standing Committee will evaluate this measure in the spring 2021 cycle.

#3623 Elective Primary Hip Arthroplasty (Acumen, LLC/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#3625 Non-Emergent Coronary Artery Bypass Graft (CABG) (Acumen, LLC/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

#3626 Lumbar Spine Fusion for Degenerative Disease, 1-3 Levels (Acumen, LLC/CMS)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the spring 2021 cycle.

Subgroup 3

During the meeting, subgroup 3 discussed three measures (#0674, #0679, and #3621) and re-voted on validity for measure #3621. To ensure reliability is consistently evaluated across the subgroups, the SMP and NQF decided to discuss any measure that has a reliability result lower than 0.50; therefore, measures #0674 and #0679 were pulled for SMP discussion. After further discussion, the subgroup accepted the preliminary analysis decision for measures #0674 and #0679. The subgroup accepted the preliminary analysis decisions for five measures (#0500, #2902, #2903, #2904, and #3501e) without further discussion. One of the nine measures evaluated by this subgroup was withdrawn from this review cycle prior to the meeting (#2689). The final results for the remaining eight measures evaluated by subgroup 3 are presented below.

#0674 Percent of Residents Experiencing One or More Falls With Major Injury (Long Stay) (Acumen, LLC/CMS)

Measure Developer/Steward Representatives at the Meeting

Cheng Lin, Sri Nagavarapu

Scientific Methods Panel Votes

- Reliability: H-0; M-6; L-2; I-0 (Pass)
- Validity: H-1; M-6; L-1; I-0 (Pass)

In their preliminary analyses, the SMP noted the developer conducted both measure score and data element reliability testing. The data element inter-rater reliability testing showed substantial agreement for gold standard nurse to gold standard nurse (kappa = 0.967) and gold standard nurse abstractor to facility nurse abstractor (kappa = 0.945) reviews. Some SMP members questioned whether the 2006-

2007 data element testing was still valid based on ongoing updates to the specifications and coding. For the measure score reliability, two tests were conducted: the signal-to-noise ratio of 0.45 suggests low/moderate reliability in separating facility characteristics from variability within the facility. The less than or equal to 0.50 reliability was the basis of the SMP discussion. The split-half Pearson's reliability (r = 0.18, ρ = 0.18, p <0.01) provided limited evidence of internal reliability as low volume "never events" contribute to small performance variations among providers. The measure received a moderate reliability rating by SMP subgroup reviewers. For validity testing, the developer also conducted measure score level in convergent validity testing with Spearman Correlations (-0.207 to +0.203) to other measures based on the hypothesis that facility percentile rankings may be somewhat consistent among similar quality measures (e.g., percent of long-stay residents who received antipsychotic, antianxiety, or hypnotic medication; percentage of long-stay residents who have depressive symptoms; and percent of long-stay residents whose need for help with daily activities has increased) with similar populations. They also hypothesized that these measures shared populations that may be likely to fall. They also conducted validity testing in variations by state, seasonality, stability, and confidence interval analyses. The empirical validity testing comparing the correlations to other measures showed weak correlations, although the demonstrated face validity was believed to be strong. SMP reviewers were concerned that 25 percent of facilities jumped three or more deciles in performance over a short interval, which may be an issue with poor reliability. One SMP reviewer stated that a fall is a never event regardless of the rate of high-risk patients and should not require risk adjustment, while another SMP member stated the absence of risk adjustment in this outcome measure "seems inconsistent with many other CMS outcome measures." After discussions, the SMP subgroup reviewers gave the measure a moderate validity rating.

Based on the similarities of the two measures, including measure developers, constructs, testing, and measure use, the SMP discussed this measure along with #0679 simultaneously. The SMP reviewers guided the discussion of whether the reliability was sufficient to warrant the moderate rating, given the marginal signal-to-noise ratio ratings and the concerns regarding the stability analysis. During the meeting, the SMP noted the data element testing was excellent with high results, although the performance score level reliability was low. Since both levels of tests are not required per NQF's current evaluation guide, the SMP decided that a moderate rating based on the data element level results is adequate. They also discussed the relatively modest signal-to-noise reliability score of 0.45, noting that the sample may have some impact on the results. The SMP also discussed the sampling of the measure with fewer than 20 stays as a restriction to the analysis when the specification determines no sampling. The SMP also questioned whether reporting performance on CMS Care Compare sites could also include the average of four rolling quarters to the four individual rolling quarters as a means to boost reliability stability. One SMP member stated that never events should put greater emphasis on the patient-level (i.e., data element) performance rather than the facility-level (i.e., measure score) performance, as the practice assumption of never events is that they should be 100 percent preventable. The SMP subgroup accepted the moderate rating in the preliminary analysis; therefore, no additional SMP vote was required. The Patient Safety Standing Committee will evaluate this measure in the spring 2021 cycle.

#0679 Percent of High-Risk Residents With Pressure Ulcers (Long Stay) (Acumen, LLC/ CMS)

Measure Developer/Steward Representatives at the Meeting

Cheng Lin, Sri Nagavarapu

Scientific Methods Panel Votes

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

In their preliminary analyses, the SMP noted the developer conducted both measure score and data element reliability testing. The data element inter-rater reliability testing showed substantial agreement for gold standard nurse to gold standard nurse (kappa = 0.92) and gold standard nurse abstractor to facility nurse abstractor (kappa = 0.97) reviews. SMP members questioned whether the 2006-2007 data element testing was still valid based on ongoing updates to the specifications and coding. For the measure score reliability, testing was assessed using a signal-to-noise ratio (0.50), suggesting moderate reliability in separating facility characteristics from variability within the facility. The less than or equal to 0.50 reliability was the basis of the SMP's discussion. The split-half reliability ($r = 0.33$, $\rho = 0.30$, $p < .01$) suggested modest evidence of internal reliability based on modest variation among providers. The SMP reviewers rated the measure's reliability as moderate. For validity testing, developers also conducted measure score level in convergent validity testing with split-half analyses of Pearson's and Spearman Correlations to other measures, hypothesizing that facility percentile rankings may be somewhat consistent among similar quality measures (e.g., *Percent of SNF Residents With Pressure Ulcers That Are New or Worsened* and Facility Five-Star Ratings). They also hypothesized that these two measures shared similar populations that may be likely to develop pressure ulcers. They also conducted validity testing in variations by state, seasonality, stability, and confidence interval analyses. The empirical validity testing comparing the correlations to other measures showed weak correlations, although the demonstrated face validity was believed to be strong. SMP reviewers were concerned that 30.4 percent of facilities jumped three or more deciles in performance over a short interval, which the reviewers reported may be attributed to low-frequency events and the impact on one event on the decile assignment. One SMP reviewer stated that a fall is a never event regardless of the rate of high-risk patients and should not require risk adjustment, while another SMP member stated the absence of risk adjustment in this outcome measure "seems inconsistent with many other CMS outcome measures". The measure received a moderate rating for validity by the SMP reviewers.

Based on the similarities between the two measures, including measure developers, constructs, testing, and measure use, the SMP discussed this measure along with #0674 simultaneously. The SMP reviewers guided the discussions of whether the reliability results were sufficient to warrant a moderate reliability. During the meeting, the SMP noted the high data element testing and moderate measure score for reliability. Since both level tests are not requested per NQF's current evaluation guide, the SMP decided that a moderate rating based on the data element level results is adequate. They also noted the age of the data element testing (2006-2007) may have potential validity effects from specification and coding updates. The SMP also discussed the sampling of the measure with greater than 20 stays as a restriction to the analysis when the measure is specified with no sampling requirements. The SMP also questioned whether reporting performance on CMS Care Compare sites could also include the average of four rolling quarters to the four individual rolling quarters as a means to boost reliability stability. One SMP member stated that never events should put greater emphasis on the patient-level (i.e., data element) performance rather than the facility-level (i.e., measure score) performance, as the practice assumption of never events is that they should be 100 percent preventable. The SMP subgroup accepted the moderate rating in the preliminary analysis; therefore, no additional SMP vote was required. The Patient Safety Standing Committee will evaluate this measure in the spring 2021 cycle.

#3621 Composite Weighted Average for 3 CT Exam Types: Overall Percent of CT Exams for Which Dose Length Product Is at or Below the Size-Specific Diagnostic Reference Level (for CT Abdomen-Pelvis With Contrast/Single Phase Scan, CT Chest Without Contrast/Single Phase Scan and CT Head/Brain Without Contrast/Single Phase Scan) (American College of Radiology)

Measure Developer/Steward Representatives at the Meeting

Karen Campos, Dustin Gress, Judy Burleson

Scientific Methods Panel Votes

- Reliability: H-5; M-2; L-0; I-1 (Pass)
- Validity: H-0; M-4; L-2; I-2 (Consensus Not Reached)
- Composite Construction: H-2; M-3; L-0; I-1 (Pass)

In their preliminary analyses, the SMP had minimal concerns with the reliability of this measure, with exceptionally high reliability (0.997) for all types of computed tomographies (CT's) and the composite weighted average in the signal-to-noise ratio analyses for physician groups and facilities greater than or equal to 10 reported patients. Data element reliability was not performed. The voting result was consensus not reached for validity during the SMP subgroup's preliminary analysis. Validity testing was conducted using face validity based on three methods: (1) a TEP of medical physicists and radiologists, (2) professional consensus recommendations, and (3) 2017-2020 use in a CMS reporting program. Some SMP reviewers were concerned that face validity was not conducted systematically by recognized independent experts and whether use in a CMS program is a true test of validity. They also questioned whether testing was based on the composite score of the individual components or the measures within the composite. The developer demonstrated that a weighted average (current measure) produces similar results to a straight average. The SMP reviewers provided a moderate rating for the composite construction.

During the meeting, the SMP discussed whether face validity testing originated from both group and facility perspectives or combined perspectives. The developers stated the information was available but not analyzed for the review. Face validity was assessed by answering "yes" to the following questions and results (out of the 21 TEP members):

1. Do you think that monitoring radiation dose indices from clinical CT exams is a good and worthwhile activity for advancing or maintaining safety and quality? (20, 95 percent)
2. Is the measure and its components as described a reasonable and appropriate way to assess performance quality of a facility or practice with regards to dose optimization? (15, 71 percent)
3. Will the scores obtained from the measure and its components as specified reasonably differentiate clinical performance across providers and separate the high performers from the low performers? (13, 62 percent)

The percent overall or average of the questions above is 48, or 76 percent. Some SMP members questioned the testing methods, stating that additional validity testing could have been conducted with the large available sample asking for split sample testing, specifically because reliability results were very high and generally higher than other composites. One SMP member stated that unusually high reliability results may indicate validity concerns. SMP members questioned the level of analysis (clinician group versus facility), specifically whether face validity was conducted at the clinician group or facility level of analysis or both levels and why stratification was conducted at the clinical group level. The developers described that the registry combines both groups and facilities. Furthermore, the developers explained that many of the providers practice in both settings and that the TIN and national provider identifier (NPI) numbers are used in the CMS reporting programs to identify reporting entities. Another SMP member questioned the accuracy of the data derived from dated scanning equipment that uses Optical Character Recognition (OCR). The developer did not have accuracy data available and acknowledged that OCR is reportedly used in low volumes and could be "finicky." SMP members also questioned the unintended consequences of low dose imaging and the need for repeat scanning. The SMP voting result was consensus not reached on validity. The Patient Safety Standing Committee will evaluate this measure in the spring 2021 cycle.

#0500 Severe Sepsis and Septic Shock: Management Bundle (Henry Ford Hospital)

Scientific Methods Panel Votes

- Reliability: H-5; M-1; L-0; I-2 (Pass)
- Validity: H-3; M-2; L-1; I-2 (Pass)
- Composite Construction: H-2; M-3; L-0; I-1 (Pass)

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

#2902 Contraceptive Care – Postpartum (HHS Office of Population Affairs)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Perinatal and Women's Health Standing Committee will evaluate this measure in the spring 2021 cycle.

#2903 Contraceptive Care – Most & Moderately Effective Methods (HHS Office of Population Affairs)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Perinatal and Women's Health Standing Committee will evaluate this measure in the spring 2021 cycle.

#2904 Contraceptive Care – Access to LARC (HHS Office of Population Affairs)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be both reliable and valid. The Perinatal and Women's Health Standing Committee will evaluate this measure in the spring 2021 cycle.

#3501e Hospital Harm – Opioid-Related Adverse Events (IMPAQ International/CMS)

Scientific Methods Panel Votes

- Reliability: H-2; M-5; L-0; I-1 (Pass)
- Validity: H-1; M-6; L-1; I-0 (Pass)

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

Discussion of Overarching Methodological Issues Identified During Measure Evaluation

Throughout the two-day web meeting, SMP members identified and discussed numerous overarching themes for the SMP to consider in upcoming advisory meetings.

SMP Methodology and Guidance

The SMP members discussed potential topics or questions for their upcoming advisory meetings to support the continued advancement of measurement science and the improvement of measure submissions, including the following:

- Offer additional validity testing education and technical assistance to developers, such as how to construct adequate tests of face validity testing, identify process measure validity comparators or correlates when correlate outcome measures are difficult to identify, and use predictive validity tests to assess whether the outcome measure is appropriate to assess performance.
- Offer proactive support of innovative development activities, both to new developers and new measure uses, especially for measures that use nonclinical conceptual frameworks and community data (e.g., community violence and food insecurity) for measurement purposes, such as measures that assess home and community-based services (HCBS), and multi-item measures with multiple performance scores.
- Review NQF's evaluation guidance for appropriateness to assess diagnostic accuracy measure constructs. Newer measures are using quality measures to assess diagnostic accuracy, which is a significant priority for the CMS Meaningful Measures Initiative. A collective discussion on the NQF evaluation criteria would assist to determine the needed content for this measure construct.
- Evaluate measured entity size, volumes, and intended use that are utilized to determine sampling methods and testing methods, as well as discussions of TINs, NPIs, incentives, and reporting based on measured entity size.

Measure Testing and Scientific Acceptability Evaluation Algorithms

The SMP members recognize the continued advancements in measurement science within the evaluation processes and asked to discuss measure testing, content in the reliability and validity algorithms, and the following in particular:

- Reliability methods and minimum acceptable thresholds recommendations for samples and volumes (especially low volumes), such as interquartiles or other descriptive statistics; additional reliability testing for small volumes; multi-year pooling for outcome measures; using process or structure measures that strongly correlate to outcomes; varying sample requirements to balance acceptable precision thresholds and intended use; volume results to assess generalizability to all hospitals based on provider characteristics, size, and rurality; and assessing whether reliability and validity can be agnostic of intended use.
- Scoring and algorithm structure challenges when the following circumstances occur: (1) accountable entity score testing is poor and the data element testing is high or moderate, (2) data element validity testing eliminates the requirement of any reliability testing, (3) data element validity testing excuses the need for data element reliability testing, and (4) data element validity "trumps" measure score reliability and validity when the priority should be on the measure score.
- Requiring performance score testing for maintenance measures for reliability and validity.
- Identifying the purpose of collapsing the high and moderate scores for reliability or validity into a single pass score.

SMP and the Consensus Development Process (CDP)

The SMP members discussed questions regarding the CDP to support the continued refinement of NQF processes, including the following:

- The improvement of carrying SMP methodologic, evaluation, and process recommendations through the CDP to the Consensus Standards Approval Committee (CSAC) and how SMP recommendations will be considered for potential modifications to evaluation guidance and the CDP
- The SMP's review challenges encountered without evidence and/or frameworks for validity assessment, especially for new measures, and diagnostic accuracy
- Whether to vote on multi-item measures (e.g., Consumer Assessment of Healthcare Providers and Systems [CAHPS] measures) with multiple performance scores individually or as a package

Public Comment

One comment was received from Don Casey, senior associate editor of the American Journal of Medical Quality, regarding the white paper that some SMP members published last year. This paper[2] provides an overview of the SMP's role in NQF's CDP and the quality measurement enterprise at large. Dr. Casey expressed interest in sharing this paper with measure developers and NQF members for a better understanding of the SMP.

Next Steps

Ms. Ingber reviewed the next steps and reminders for the SMP and 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 spring 2021 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) 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 are not provided or otherwise did not meet NQF's minimum evaluation requirements. Measures #2881 and #3188 are eligible for Standing Committee discussion and a revote, if pulled by their respective Standing Committees. Endorsement will be removed for these two maintenance measures that did not pass the SMP's vote if they are not selected for further discussion and a revote. Measures moving forward in the spring 2021 evaluation cycle will be reviewed by their respective Standing Committees in June or July 2021 and discussed by the CSAC in November 2021.

The SMP will convene via web meeting on May 4, 2021, to continue a discussion on acceptable reliability thresholds and risk adjustment. Additionally, SMP members were invited to join a web meeting with the NQF-convened TEP on Best Practices for Developing and Testing Risk Adjustment Models on May 13, 2021 from 1:00pm – 3:00pm ET.

[2] Nerenz DR, Cella D, Fabian L, et al. The NQF Scientific Methods Panel. *Am J Med Qual*. 2020;35(6):458-464.