NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel Measure Evaluation Web Meeting- Spring 2020

The National Quality Forum (NQF) convened the Scientific Methods Panel (SMP) on April 1-2, 2020 for a discussion of the scientific properties (reliability and validity) of several complex measures submitted to the Spring 2020 evaluation cycle. Of the 21 measures reviewed by the SMP this cycle, seven measures were discussed during this meeting, including those for which subgroup members did not reach consensus in their preliminary evaluations, those that did not initially pass the SMP evaluation but for which measure developers provided additional information, and those that were otherwise pulled for discussion by NQF staff or subgroup members. The SMP also discussed several overarching methodological issues which were identified based on their review of the measures this cycle. A brief summary of the discussion of these issues, the seven measures discussed during the web meeting, and voting results for the measures not discussed during the webinar are included in this document.

Welcome, Introductions, and Review of Meeting Objectives

Ashlie Wilbon, NQF Senior Technical Expert, welcomed the members of the Panel and participants to the web meeting. NQF's CEO, Shantanu Agrawal, and Scientific Methods Panel co-chairs David Nerenz and David Cella, also provided opening remarks. Ms. Wilbon asked Panel members to introduce themselves and provide any disclosures of interest relevant to the measures to be discussed during the meeting. Ms. Wilbon then described the process for the measure discussions and reviewed relevant NQF evaluation criteria.

Discussion of Overarching Methodological Issues Identified During Measure Evaluation

The SMP identified several overarching methodological issues for discussion that arose over the course of their measure evaluations this cycle. The Panel agreed that these initial discussions would serve as the foundation for future papers to be authored by SMP members and NQF to help support recommendations to NQF on its policies and criteria, as well as provide guidance for measure developers on these challenging methodologic issues. A brief summary of each of these issues is provided below:

Reliability

For several cycles, the Panel has recognized the challenges regarding the lack of consensus on acceptable thresholds for measure score reliability statistics. This cycle there were several measures that highlighted this challenge once again. Multiple sources of literature have been referenced by measure developers, panel members, and NQF in its 2011 testing task force report, which suggested thresholds for acceptable reliability scores (e.g., Landis, et al, and Adams, et al). However, the panel recognized that a critical review of the literature and the context in which these thresholds are applied is needed. Moreover, the Panel agreed that a single threshold that would be applicable for all reliability statistics is not feasible. The Panel also recognized that the evaluation of reliability, including the methodology and interpretation of results, should be done in the context of how the measure will be used. For example, a lower threshold for a particular statistic may be acceptable if a measure will be

used for quality improvement, as opposed to a pay-for-performance program. In other cases, the reliability testing approach employed may only demonstrate reliability for a particular application (e.g., identification of outliers). The panel raised several important challenges presented by this. While NQF considers use in the recommendation for endorsement, it is typically not incorporated into the review of scientific acceptability and endorsement is granted agnostic to a specific use; rather, NQF's current process grants endorsement and signals the measure is appropriate for use in any accountability application. NQF is currently exploring these issues and expects further discussion of this ongoing challenge with the Consensus Standards Approval Committee (CSAC). The Panel will also continue to discuss these issues during upcoming webinars.

The Panel also expressed the importance of identifying minimum sample sizes in the testing analyses and then discussed the relationship between reliability and validity. More specifically, they agreed that reliability can be impacted by the adequacy or inadequacy of the risk adjustment model and that this should be a consideration when assessing reliability.

Social Risk Adjustment

Current guidance for the SMP's evaluation of the risk strategy states that they should not vote "Low" on validity (therefore not passing the measure), solely due to a developer's decision not to include (or to include) social factors in the risk adjustment model. The SMP is asked to focus their evaluation on the calibration, discrimination, and testing of the model. However, the panel articulated the challenge this poses as the factors included in the model, both clinical and social, are integral to the calibration and discrimination of a model. Another issue identified by the Panel was the concern that some developers' decision not to include social factors in their risk model is not supported by the empirical and conceptual analyses or the recommendations laid out in the 2014 NQF social risk adjustment paper. The Panel discussed ways in which they may be able to communicate and signal their concerns regarding social risk adjustment in a measure to the relevant standing committee beyond the current process which focuses on relaying qualitative feedback.

The Panel and NQF staff discussed the current process for the panel's evaluation and the standing committees' discussions of risk adjustment and how this process could be improved to ensure this information is appropriately communicated. Suggestions including potentially enabling the SMP to fully consider risk adjustment in their evaluation of validity and still allowing the standing committees to re-vote validity if a measure were to not pass for this reason. Other suggestions focused on ensuring the Committee are well-informed of the guidance supporting the social risk trial in order to facilitate consistent evaluation of the risk model.

Cost measure evaluation challenges

There were several cost measures submitted for evaluation this cycle, all of which passed reliability and validity. However, there were some questions raised regarding various methodological approaches employed in the measures including the risk adjustment approach and details of measure exclusions.

Reviewers questioned whether the risk adjustment approach was tailored enough to the specific measure focus beyond the standard Hierarchical Condition Category (HCC) model and whether sufficient detail was provided for how exclusions were established. Some Panel members commented that the HCC risk model is well-validated for use in cost measures and should be an acceptable approach to adjusting for risk. The Panel also agreed that additional details on the rationale for certain exclusions would be helpful to better understand the validity of the measure. The Panel ultimately determined that

based on current requirements these issues were adequately addressed by developers. Some panel members who also recently attended a webinar convening the NQF Cost and Efficiency Standing Committee, shared some of the challenges discussed regarding evaluating validity. These challenges included the selection of a comparator to demonstrate construct or criterion validity and evaluating validity in the context of how it will be used. A more detailed examination of these issues and other challenges with the evaluation of cost and resource use measures will be addressed in a future paper by panel members.

Measure Evaluation

Measure evaluations during the webinar 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 seven of the 21 measures under consideration this cycle. Subgroup members then performed in-depth reviews and analyses of their assigned measures. Developers received these preliminary analyses prior the meeting and were given an opportunity to submit written responses to concerns expressed in the analyses. These responses were provided to the SMP prior to the meeting for review to support their discussion and subsequent voting on the measures during the meeting.

During the meeting, the SMP evaluated reliability and validity for seven measures based on their preliminary analyses and additional information submitted for consideration by the developer. 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. David Cella and David Nerenz, SMP co-chairs, facilitated the remainder of the discussion, wherein a lead discussant from the subgroup that first evaluated the measure noted the primary concerns of the subgroup. Other subgroup members made additional comments. The SMP co- chairs then 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. The subgroup members who provided an in-depth preliminary analysis of the measure voted on the measure then submitted final votes for the relevant criteria. Quorum was achieved for all subgroup votes. These votes reflect the final overall assessment of reliability and/or validity by the SMP.

The remaining 14 of the 21 measures evaluated by the SMP in the Spring 2020 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 will serve as the final overall assessment of reliability and validity for the standing committees' consideration.

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

Subgroup 1

Subgroup 1 discussed four measures (3559, 3556, 0715 and 3576) and accepted the preliminary analysis decisions for three measures (0076, 0716, and 2687) without further discussion. The results for the seven measures evaluated by Subgroup 1 are presented below.

3559 Hospital-Level, Risk-Standardized Improvement Rate in Patient-Reported Outcomes Following Elective Primary Total Hip and/or Total Knee Arthroplasty (THA/TKA) (CMS/Yale/YNHH Center for Outcomes Research and Evaluation (CORE)

Measure Steward/Developer Representatives at the Meeting

Lisa Suter, Kathleen Balestracci, Darinka Djordjevic, Victoria Taiwo

Scientific Methods Panel Votes

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

In their preliminary analyses, the subgroup members found the measure to be reliable, but consensus was not reached on validity. Reviewers identified several concerns related to missing data, exclusions, and the attribution approach. One panel member also raised concern regarding the impact of this measure given the selection of outcome measures, HOOS, JR and KOOS, JR, on the measurement landscape, alignment with registries, and other similar approaches. Developers provided a detailed response to the reviewers' concerns on these issues including a summary of the development process which relied heavily on technical experts and patients in particular; this process guided the selection of the patient reported outcome measure/instrument. The developers clarified the rationale for minimum case size of 25 per hospital, and the exclusions of staged procedures. The developers also noted support for this measure among orthopedic societies. After weighing these concerns with the developers' responses, the panel passed the measure on validity and the Patient Experience and Function Standing Committee will evaluate this new measure in the Spring 2020 cycle.

3556 National Healthcare Safety Network (HNSN) Nursing Home-Onset Clostridiodes difficile Infection (CDI) Outcome Measure (Centers for Disease Control and Prevention)

Measure Steward/Developer Representatives at the Meeting

Jeneita Bell, Jonathan Edwards, Elizabeth Mungai, Suparna Bagchi

Scientific Methods Panel Votes

- Reliability: H-0; M-0; L-8; I-1 (Not Pass)
- Validity: H-0; M-0; L-8; I-1 (Not Pass)

In their preliminary analyses, subgroup members identified several concerns with the data element validity testing submitted for this measure. NQF criteria does not require reliability testing be submitted if data element validity testing has been performed; when this policy is applied, the vote for validity also serves as the vote for reliability as was the case for this measure. The concerns regarding the data element validity testing consisted of a lack of patient-level factors included in the risk model, variation in the validation process among states reported in the testing data, and a lack of testing for all critical data elements. During the discussion of the measure during the meeting, the developer team responded to these concerns noting a lack of patient-level data and varied state level reporting requirements as challenges addressing these concerns. Ultimately, the panel voted not to pass the measure given their lingering concerns. This new measure may be resubmitted in a future cycle for reconsideration by the SMP at the developer's discretion.

0715 Standardized Adverse Event Ratio for Children <18 Years of Age Undergoing Cardiac Catheterization (Boston Children's Hospital – Center for Excellence for Pediatric Quality Measurement)

Measure Steward/Developer Representatives at the Meeting

Lisa Bergersen

Scientific Methods Panel Votes

- Reliability: H-0; M-8; L-0; I-0 (Pass)
- Validity: H-0; M-3; L-1; I-4 (Not Pass)

In their preliminary analyses, the subgroup did not pass this measure on reliability and validity due to several concerns. For the reliability testing, reviewers expressed concern regarding the adequacy of the statistical analysis to demonstrate data element reliability, noting the lack of representativeness of the sample and lack of testing for all critical data elements. For the validity testing, the subgroup expressed concerns with the methodology used and questioned its ability to demonstrate validity of the measure score. The developer provided a detailed response to the subgroup's concerns and modified the testing approach for both reliability and validity, which addressed the reviewer's concerns with reliability. The developer also submitted revised validity testing, noting several revisions to the measure since its last endorsement. However, this testing focused on the risk model rather than the measure score or data elements and therefore did not pass the validity criterion.

3576 Pediatric Asthma Emergency Department Use (University of California, San Francisco)

Measure Steward/Developer Representatives at the Meeting

Naomi Bardach

Scientific Methods Panel Votes

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

In their preliminary analyses, the subgroup did not pass this measure on reliability and validity due to several concerns. For the reliability testing, the testing did not appear to align with the level of analysis (health plan) for which the measure was specified, and the testing approach and results presented very low intraclass correlation statistics (ICC). For the validity testing, the subgroup expressed concerns with the methodology used and questioned its demonstration of validity of the measure score by examining the impact of the measure in a quality improvement initiative. The developer provided a detailed response to the subgroup's concerns and significantly modified the testing approach for both reliability and validity. While the revised reliability testing approach addressed some of the reviewer's concerns, and presented a higher ICC, there were still lingering concerns. The developer also submitted revised validity testing. However, this testing focused on the risk model rather than the measure score or data elements and therefore did not pass the validity criterion. This new measure may be resubmitted in a future cycle for reconsideration by the SMP at the developer's discretion.

0076 Optimal Vascular Care (MN Community Measurement)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cardiovascular Standing Committee will evaluate this measure in the Spring 2020 cycle.

0716 Unexpected Complications in Term Newborns (California Maternal Quality Care Collaborative)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Perinatal and Women's Health Standing Committee will evaluate this measure in the Spring 2020 cycle.

2687 Hospital Visits After Hospital Outpatient Surgery (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Surgery Standing Committee will evaluate this measure in the Spring 2020 cycle.

Subgroup 2

During the meeting, the subgroup discussed one measure (2496). The subgroup accepted the preliminary analysis decisions for six measures (3561, 3562, 3563, 3564, 3574, and 3575) without further discussion. The final results for the seven measures evaluated by subgroup 2 are presented below.

2496 Standardized Readmission Ratio (SRR) for Dialysis Facilities (The Centers for Medicare and Medicaid Services)

Measure Steward/Developer Representatives at the Meeting

Casey Parrotte, Joe Messana, Jesse L. Roach, Joel Andress, Wilfred Agbenyikey, Jennifer Sardone, Jack Kalbfleisch, Claudia Dahlerus

Scientific Methods Panel Votes

- Reliability: Consensus not reached
- Validity: H-0; M-3; L-5; I-0 (Not Pass)

In their preliminary analyses, subgroup reviewers did not pass this measure on validity and consensus was not reached on reliability. Reviewers raised concerns with the measure score reliability testing result, which was considered modest/low. Given the similar methodology used in testing score-level reliability between this measure and others from the same developer reviewed this cycle, the panel ultimately determined that consensus could not be reached on the reliability without established guidance on thresholds for reliability testing; no subgroup vote was recorded for this vote. It was therefore decided that the final vote on reliability should lie with the Standing Committee, which will evaluate all of the measures with similar methodologies together and determine the adequacy of the results across all similar measures to demonstrate reliability. For validity, the concerns centered on the adequacy of the correlations presented for measure score validity testing. The developers provided a detailed response to the panel's concerns. However, reviewers still found the results did not adequately demonstrate measure score validity and did not pass the measure on validity. NQF's most recent policy on measures that will be eligible for review by standing committees following SMP review states that measures that did not pass for a reason other than inappropriate methodology or inadequate testing, can be reconsidered and voted upon by the standing committee if the committee chooses to do so. Therefore, this measure will be eligible for consideration and re-vote by the Admissions and Readmissions Standing Committee in the Spring 2020 cycle.

3561 Medicare Spending Per Beneficiary – Post Acute Care Measure for Inpatient Rehabilitation Facilities (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

- Reliability: H-3; M-4; L-0; I-0 (Pass)
- Validity: H-1; M-6; L-1; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

3562 Medicare Spending Per Beneficiary – Post Acute Care Measure for Long-Term Care Hospitals

(The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

3563 Medicare Spending Per Beneficiary – Post Acute Care Measure for Skilled Nursing Facilities (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

3564 Medicare Spending Per Beneficiary – Post Acute Care Measure for Home Health Agencies (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

3574 Medicare Spending Per Beneficiary (MSPB) Clinician (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

3575 Total Per Capita Cost (TPCC) (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Cost and Efficiency Standing Committee will evaluate this measure in the Spring 2020 cycle.

Subgroup 3

Subgroup 3 discussed two measures (2539 and 3566) during the meeting and accepted the preliminary analyses decisions for five measures (0369, 1463, 2977, 2978, and 3565) without further deliberation. The final results for the seven measures evaluated by subgroup 3 are presented below.

2539 Facility 7-Day Risk-Standardized Hospital Visit Rate After Outpatient Colonoscopy (The Centers for Medicare and Medicaid Services)

Measure Steward/Developer Representatives at the Meeting

Doris Peter, Elizabeth Drye, Craig Parzynski

Scientific Methods Panel Votes

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

In their preliminary analyses, the subgroup passed the measure on reliability; however, consensus was not reached on validity. The subgroup primarily raised concern with the developer's rationale for not providing empirical analyses of the validity testing for maintenance review. The developer provided a detailed written and verbal response to these concerns which facilitated a discussion with the panel on the other types of validity testing that could have been conducted other than what was described by the developers, and the feasibility of those testing approaches. The developer was amenable to exploring other types of validity testing, but raised concerns with the feasibility of performing additional testing as well as whether the alternative methods suggested would meaningfully demonstrate validity of the measure score. Given the lack of a clear alternative for validity testing, the panel voted to pass the measure on validity and this measure will be considered by the Admissions and Readmissions Standing Committee for the Spring 2020 cycle.

3566 Standardized Ratio of Emergency Department Encounters Occurring Within 30 Days of Hospital Discharge (ED30) for Dialysis Facilities (UM – Kidney Epidemiology and Cost Center)

Measure Steward/Developer Representatives at the Meeting

Casey Parrotte, Joe Messana, Jesse L. Roach, Joel Andress, Wilfred Agbenyikey, Jennifer Sardone

Scientific Methods Panel Votes

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

In their preliminary analyses, subgroup reviewers did not pass this measure on reliability and passed the measure on validity. Reviewers raised concerns with the measure score reliability testing result, which was considered modest/low and questioned whether the testing approach used, namely the provider inter-unit reliability (PIUR), was intended to demonstrate the measure was reliable only for the purpose of identifying outliers. Given the similar methodology used in testing score-level reliability between this measure and others reviewed from the same developer this cycle, the panel ultimately determined that consensus could not be reached on the reliability without existing guidance on thresholds for reliability testing; no subgroup vote was recorded for this vote. It was therefore decided that this evaluation should be left to the Standing Committee, which will evaluate all of the measures with similar methodologies together and determine the adequacy of the results across all similar measures to demonstrate reliability. Therefore, this measure will be eligible for consideration and re-vote by the Admissions and Readmissions Standing Committee in the Spring 2020 cycle.

0369 Standardized Mortality Ratio for Dialysis Facilities (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Renal Standing Committee will evaluate this measure in the Spring 2020 cycle.

1463 Standardized Hospitalization Ratio for Dialysis Facilities (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

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

Subgroup members found the measure to be reliable and valid. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the Spring 2020 cycle.

2977 Hemodialysis Vascular Access: Standardized Fistula Rate (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

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

Subgroup members found the measure to be reliable and valid. The Renal Standing Committee will evaluate this measure in the Spring 2020 cycle.

2978 Hemodialysis Vascular Access: Long-Term Catheter Rate (The Centers for Medicare and Medicaid Services)

Scientific Methods Panel Votes

- Reliability: H-4; M-5; L-0; I-0 (Pass)
- Validity: H-1; M-6; L-2; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The Renal Standing Committee will evaluate this measure in the Spring 2020 cycle.

3565 Standardized Emergency Department Encounter Ratio (SEDR) for Dialysis Facilities (UM – Kidney Epidemiology and Cost Center)

Scientific Methods Panel Votes

- Reliability: H-2; M-6; L-1; I-0 (Pass)
- Validity: H-1; M-5; L-3; I-0 (Pass)

Subgroup members found the measure to be reliable and valid. The All-Cause Admissions and Readmissions Standing Committee will evaluate this measure in the Spring 2020 cycle.

Public Comment

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

Next Steps

The NQF Scientific Methods Panel team will inform developers and standing committees of the SMP discussion and votes. Measures that passed for both reliability and validity or were consensus not reached will be considered by the relevant standing committees in the Spring 2020 evaluation cycle. Measures that did not pass the SMP vote may be pulled for discussion by the relevant standing committee. Of the measures that did not pass, measure #2496 will be eligible for standing committee re-vote if they choose. The remaining measures that did not pass (3556, 3576, 0715) will not be eligible for re-vote and may be resubmitted to a future cycle; endorsement may be removed for maintenance measures that did not pass.

The SMP will reconvene via webinar on May 26, 2020 to discuss recommendations for modifying the current NQF criteria and guidance for reliability and validity testing.