NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel July Web Meeting

The National Quality Forum (NQF) convened a public web meeting for the Scientific Methods Panel (SMP) on July 29, 2021.

Welcome, Introductions, and Review of Web Meeting Objectives

Tricia Elliott, NQF senior managing director, began by welcoming participants to the web meeting. SMP Co-Chairs Drs. David Nerenz and Christie Teigland also provided opening remarks. Ms. Elliott reviewed the following meeting objectives: (1) consider a policy change that requires validity testing at the accountable-entity (formerly known as measure score) level for all maintenance measures, (2) review forthcoming updates to the measure evaluation guidance, and (3) improve and clarify guidance on reliability testing for future measure development and evaluation cycles.

Requiring Validity Testing at the Accountable Entity Level for Maintenance Measures

Dr. Teigland led the discussion that proposed requiring accountable-entity level validity testing for all maintenance measures. During the previous SMP advisory meeting on May 4, 2021, and in multiple subsequent SMP advisory calls and measure evaluation review meetings, SMP members requested formal consideration for this requirement. SMP members also raised concern with NQF's current policy that allows the SMP to pass measures submitted for maintenance evaluation with only patient-/encounter-level testing, which may deprioritize performance assessment at the accountable-entity level. Dr. Teigland summarized the previous discussions. Reasons for favoring this policy change include NQF's *use* subcriterion (i.e., requiring performance results in at least one accountability application within three years after initial endorsement for maintenance measures), which provides accountable-entity level data for validity testing. For reasons not in favor of this policy change, Dr. Teigland asked SMP members for examples in which only patient-/encounter-level testing would be prioritized at maintenance. Some individuals noted that empirical validity testing with process measures may be difficult if no other gold standard measures correlate for empirical validity testing. In this case, patient/encounter testing might be the only option available. Other SMP members generally agreed with the recommendation, yet exceptions would be considered when a strong rationale is provided. The SMP members agreed that NQF's maintenance validity testing should require accountable-entity level testing.

As with other panel discussions, SMP members also reiterated that NQF guidance should discourage measure validity testing between measures that are essentially autocorrelated. The rationale for choosing measures for empirical validity testing should reflect a conceptual model that demonstrates a process-outcome relationship. SMP members acknowledged rare instances of testing correlations in the opposite direction, but this is not preferred. Dr. Nerenz stated that the SMP co-chairs provided similar guidance to developers at the Measure Developer Workshop on June 7, 2021.

Following a robust discussion, the SMP made two recommendations for validity testing in maintenance evaluations:

1. Empirical validity testing at the accountable-entity level should be required for all maintenance measures. If measure developers are unable to meet this requirement, NQF should require, and developers should provide, a strong rationale supporting this rare instance.
2. Measures submitted for maintenance evaluations with face validity testing should include other level validity testing (i.e., accountable-entity validity testing or, in rare instances, patient-/encounter-level validity testing with a strong rationale for not performing accountable-entity validity testing).

The SMP members also agreed that further discussion was needed regarding the following items: (1) acceptable rationales for not performing accountable entity validity testing and (2) prioritizing accountable-entity level and empirical validity testing for maintenance measures when patient-/encounter-level testing is also submitted. These concepts will be discussed during future SMP advisory meetings.

Measure Evaluation Guidance Updates

Ms. Elliott provided an overview of measure evaluation guidance updates, which are anticipated to be published by NQF in August 2021. Each year, NQF revisits this document to ensure that the most current guidance is available to measure developers. This year, updates include clarifications on language, definitions, and more detailed guidance for specific measure types. Specifically, NQF added definitions of *patient-reported outcomes*, *patient-reported outcome measures*, and *patient-reported outcome performance measures*. To comport with the SMP's latest recommended language, NQF also updated all language from *data element* to *patient-/encounter-level* and from *measure score* to *accountable-entity level*. There are certain areas of the guidance in which the phrase *data element* is retained, as it can sometimes refer to how a measure is constructed, especially with electronic clinical quality measures (eCQMs). Lastly, NQF added additional testing guidance for composite measures.

Reliability Testing Thresholds

Dr. Sharon Hibay, NQF senior consultant, introduced the next discussion item: continued discussions on the reliability testing thresholds table. Dr. Hibay reminded participants that this table was introduced during the last SMP web meeting and is meant to serve as a guiding tool for reliability testing for the purpose of NQF endorsement, as different testing approaches have different thresholds for reliability. Updates were made to the table since the last meeting to reflect that discussion. Only a minimum reliability threshold is included and some testing types were eliminated to remove duplications and improve clarity. The SMP was reminded that the revised reliability table was presented to guide the panel discussion but was not presented as a final document. Dr. Hibay also discussed the importance of vetting reliability testing threshold recommendations throughout the measurement community, including, but not limited to, an upcoming public commenting period to seek NQF member, measure developer, and public feedback on any draft recommendations to have a greater understanding of the effect of these potential changes. She also discussed that any recommendations would require approval of the Consensus Standards Approval Committee (CSAC) and become operationalized by NQF staff. The SMP webpage has a copy of the discussed table: DRAFT Acceptable Reliability Thresholds (Version 3.02).

Dr. Nerenz guided the discussion and noted a subgroup of SMP members continued their robust dialogue for the document revisions. He and other SMP members encouraged active discussion participation of all Panel members, with a goal of defining a threshold by the end of the meeting. He asked the SMP members to start with accountable-entity level reliability testing because they previously agreed that this level of testing should take precedence over patient-/encounter-level testing. Without objection, the SMP agreed and reviewed the two accountable-entity reliability testing approaches: (1)

Signal-to-Noise Ratio (SNR) or Inter-Unit Reliability (IUR) and (2) Split-half reliability testing (e.g., intraclass coefficient, with correction for full sample with the Spearman-Brown formula). Dr. Nerenz briefly discussed that the deleted test-retest reliability approaches may require a separate table than the patient-/encounter-level and accountable-entity reliability testing. For each testing level, a testing approach, purpose, range, and threshold were provided. SMP Member Dr. Larry Glance noted that the range for both approaches should be 0 to 1 rather than -1 to 1, which was depicted in the SNR approach. NQF staff will make that correction. The threshold of 0.5 was proposed to initiate Panel discussion for both testing approaches, which essentially states that approximately half of the overall observed variance is signal, and the other half is noise. Dr. Nerenz stated that the basis of the 0.5 selection was based on extensive literature, measures, practical expertise, and extensive discussion of the subgroup while balancing with the current 0.4 assumed threshold. Dr. Nerenz previously discussed that the original value of 0.4 came from the 1977 Landis and Koch article titled "The Measurement of Observer Agreement for Categorical Data", which was for Cohen's kappa statistic for the gold standard interrater reliability reviews (IRRs) that became the "assumed" threshold for all reliability testing. One SMP member stated that during the initial meeting of the SMP in 2017, the concept of increasing the assumed threshold was discussed with a general agreement that 0.4 was too low. No comments were received from any SMP members to continue with this current threshold.

Multiple SMP members spoke in favor of 0.6 as a minimum threshold, although numerous credible and trusted studies prescribe 0.7 as a minimum threshold. A few SMP members stated that the difference between 0.4, the current assumed minimal threshold, and 0.7 and 0.8 may be too wide of a gap and may negatively eliminate many measures in the portfolio. Other SMP members expressed concern with the drive for high reliability (i.e., 0.6 and higher), which could remove measures with lower reliability that still drive some improvement. One SMP member stated that an overly elevated minimum reliability threshold may impede providers, medical groups, and hospitals with wide variations in volume, specifically low-volume providers in which measure reliability may be lower. Other SMP members were concerned that low-volume providers could be given a "pass" on reporting and/or performing. One SMP member stated their experience with post-acute and skilled nursing facility settings demonstrates high-provider volumes not included in performance due to low volumes. Another SMP member stated that reliability is generally reported in a mean or median of distribution. Therefore, with a minimum threshold of 0.5, the left tail of reliability would be between 0.1 and 0.2, which is an inherent function for volume distributions.

In response, another SMP member stated that reliability thresholds and classification stability have limited definitive mapping to each other, yet the combination explain information about confidence intervals. Some SMP members recommended a phased approach (e.g., 0.5 for a period of time and then 0.6) in a conservative approach to elevating the threshold, while others suggested weighting measures based on reliability results. Others stated that the recommendation would not be permanent and could be adjusted after implementation as needed. Other SMP members questioned how a threshold would be operationalized through the Consensus Development Process (CDP) projects, which NQF staff stated would be guided by significant vetting throughout the measurement community. A few other SMP members stated that the role of the SMP was to set a higher scientific standard, and the SMP and CDP Standing Committees would assimilate the recommendations during measure evaluations. Some SMP members also wanted to understand the impact across measures in CDP portfolios, reliability, and by volumes. One SMP member suggested that an analysis could be conducted prior to finalizing an SMP recommendation. With perceived greater support for the 0.6 threshold, one SMP member proposed conducting an informal straw poll to vet 0.6 as the minimum reliability threshold and allowing providers to submit measures at their risk with a rationale for accepting a reliability below 0.6. This proposal was generally accepted by SMP members, although not in total. Straw poll voting was not officially tallied.

Public Comment

Dr. Hibay opened the web meeting to allow for public comment. One commenter asked for confirmation of whether the question regarding thresholds for SNR referred to the mean, median, minimum, or some percentile of the distribution. Dr. Nerenz explained that the discussion of the table was flexible around this expression of central tendency to ensure it applies across the range of measures the SMP sees. However, once the table is finalized, accompanying language will make this matter clearer.

Another commenter requested clarification on the SMP's working definition of *reliability*. Dr. Nerenz explained that the SMP uses multiple concepts of reliability (e.g., temporal stability, precision of measurement, and stability misclassifications), recognizing that this presents some difficulty to developers. However, it also allows the SMP to consider measures from multiple perspectives.

Next Steps

The SMP will confirm their positions on the SNR/IUR threshold and person-/encounter-level half of the table offline. NQF staff will prepare the table for public comment and review by the CSAC. The next SMP web meeting will be the fall 2021 measure evaluation meeting, which will take place on October 26–27, 2021.