NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel December Web Meeting

The National Quality Forum (NQF) convened a public web meeting for the Scientific Methods Panel (SMP) on December 14, 2021.

Welcome, Introductions, and Review of Web Meeting Objectives

Kathleen Giblin, NQF senior vice president, welcomed participants to the web meeting. SMP Co-Chairs Drs. David Nerenz and Christie Teigland also provided opening remarks. Ms. Giblin reviewed the following meeting objectives: (1) Discuss feedback on minimum reliability testing thresholds from the Consensus Standards Approval Committee (CSAC), the Measure Developer Advisory Panel, and the Standing Committee Advisory Group; (2) Discuss feedback on accountable entity level reliability and validity testing requirements from the CSAC, the Measure Developer Advisory Panel, and the Standing Committee Advisory Group; (3) Confirm consensus on recommendations of minimum reliability testing thresholds for patient/encounter and accountable entity levels; and (4) Confirm consensus on recommendations for reliability and validity testing at the accountable entity level for maintenance measures.

Confirm Consensus on Recommendations for Minimum Reliability Testing Thresholds

Hannah Ingber, NQF senior analyst, introduced the first item for discussion: confirming the SMP's consensus on the reliability testing thresholds table following its review by the CSAC, the Measure Developer Advisory Panel, and the Standing Committee Advisory Group. Ms. Ingber presented feedback, questions, and recommendations from the varying groups. All three groups expressed gratitude to the SMP for addressing this issue and generally agreed with the proposed thresholds, except for the inter-rater agreement threshold within person/encounter level reliability testing. All three groups proposed that the SMP consider a threshold above 0.4. Additionally, the CSAC presented recommendations for the final guidance document that the SMP will produce, which included the following items: (1) guiding principles for the threshold application, (2) next steps for measures that do/do not meet testing minimums, and (3) rationale and cited references for each testing threshold. Ms. Ingber then presented feedback from the Measure Developer Advisory Panel, which requested greater detail and guidance on testing requirements and methods specific to the data collection versus the measure type, as well as the patient-reported outcome performance measures (PRO-PMs) versus clinician-reported/-interpreted measures. Additionally, the Measure Developer Advisory Panel members requested statements clarifying potential exceptions to the guidance, such as measures that target low-volume, high-priority "never" events, supporting literature explaining low testing results, and other testing requirement constraints by measure or data collection types. Lastly, Ms. Ingber presented feedback from the Standing Committee Advisory Group, which expressed the need for a well-planned rollout of the thresholds to ensure consistent Consensus Development Process (CDP) Standing Committee measure evaluation. The Advisory Group recommended that successful implementation include robust documented guidance and training for all stakeholders that include the following components: (1) the purpose and use of the threshold tables to serve as a discussion guide for the Standing Committee

evaluations; (2) the SMP's decision making framework, supporting references, refuting rationales, and justifications for not meeting the thresholds; and (3) a supplement of graduated testing ranges below, above, and near the thresholds to guide the interpretation of results. Following the overview of recommendations from the three groups, Ms. Ingber turned the meeting over to Dr. Nerenz to lead the SMP in further discussion.

Dr. Nerenz restated the SMP's rationale for proposing the initial inter-rater agreement threshold of 0.4, citing the 1977 Landis and Koch article titled "The Measurement of Observer Agreement for Categorical Data." Dr. Nerenz elaborated further on the impact of the achievable value of kappa and its dependence on the underlying rate of the recorded events. This article supported the SMP's decision to propose the 0.4 threshold initially. Dr. Nerenz then cited the article titled "Interrater reliability: the kappa statistic" written by Mary McHugh in support of a 0.6 threshold. The article provides a rationale for increase by using the concept of error (i.e., what fraction of the data elements should be presumed to be incorrect). Using this method, one can determine what would be an acceptable level of error and thereby establish a threshold. McHugh argues that a minimum threshold of 0.6 must be achieved to be confident that at least half the data elements are accurate (i.e., this should be the lowest threshold to ensure reliability.)

Dr. Nerenz solicited additional input on the increased threshold from the SMP. None of the SMP members argued against raising the threshold; rather, the conversation focused on the exceptions to this threshold and how those exceptions might be applied within NQF's measure evaluation process. The main exceptions that were discussed were low or high prevalence events and the use of the measure. To address these exceptions, SMP members recommended that each line of the threshold tables includes guidance or explanatory text (i.e., for developers and the SMP) on the referenced statistical methods and most appropriate uses. Additionally, the SMP suggested that this guidance include considerations for exceptions and acceptable justification/rationale for not meeting the threshold. Justifications discussed included the use of the maximum achievable kappa used as a comparator to the kappa that the developer calculated, using confidence intervals to capture the susceptibility of the kappa to low-prevalence conditions, and presenting data on how the measure would be applied in use and whether it is still producing sufficiently reliable information on comparative performance. The SMP further suggested that if the stated exceptions include a statistic that they believe would be better suited to accurately represent the data, the SMP would make those recommendations within this explanatory text. One SMP member noted the "use-agnostic view" that the SMP is encouraged to consider when evaluating measures and further noted that the conversation surrounding kappa is about tolerance for error, which can change under differing circumstances and the use of the measure. This SMP member further stated that as the SMP continues to evolve their guidance surrounding these thresholds, NQF should continue to assess the parameters of measure use (e.g., use-agnostic versus Use-specific) within the context of the measure evaluation criteria.

Dr. Nerenz asked NQF staff about the use of the tables and set of standards and how these new thresholds would apply to the overall endorsement process, specifically asking for staff to clarify the degree of flexibility that NQF, the SMP, and Standing Committees should use when applying the thresholds. Ms. Ingber stated that NQF is looking to take a softer approach to the implementation of the thresholds and welcomes the SMP's feedback and suggestions. An SMP member stated that they are not advocating for a strict threshold, explaining that if a developer submits a measure with values lower than the prescribed thresholds, the SMP should always discuss and evaluate the submission. In order for a proposed measure to receive NQF endorsement, the measure should have reliability at or above these recommended thresholds, as demonstrated in appropriately selected statistical tests. These are the standards that measures are expected to meet under typical testing situations; however, the developer

will need to provide a detailed rationale for not meeting the thresholds. There were no objections to this statement from other SMP members, signaling general agreement with the statement.

Dr. Nerenz closed this portion of the meeting by stating that the SMP will move forward with the 0.6 threshold for inter-rater agreement reliability at the person/encounter level. Furthermore, Drs. Nerenz and Teigland and NQF staff will begin drafting the explanatory notes that will accompany the table and forward them to the SMP members for comment.

Confirm Consensus on Recommendations for Accountable Entity Level Reliability and Validity Testing Requirements for Maintenance Measures

Reliability Testing Requirements

Ms. Ingber introduced the next item of discussion: confirming the SMP's consensus on their prior recommendation to require accountable entity level reliability testing at maintenance. This item was also reviewed by the Advisory Groups. NQF socialized the information with these groups, and all agreed that this should be required; however, they asked for additional consideration of three potential exceptions to this rule, as noted below. Dr. Teigland led the discussion.

The other Advisory Groups raised concerns about data collection for testing at the accountable entity level within the three-year maintenance cycle. Some SMP members were skeptical about this assertion, noting that at maintenance, the measure should be in use for a few years. However, others articulated several reasons for a delay in data availability, even within three years. These data restrictions are due to external administrative/structural reasons, including contract changes between measure developers, federal agency priorities, and regulatory timelines. This can be particularly challenging for electronic clinical quality measure (eCQM) testing, which necessitates large samples from many partner hospitals. Considering these challenges, the SMP discussed the possibility that measure maintenance may be delayed beyond three years to accommodate these situations. These justifications would be taken into consideration under the current three-year review period. However, under normal circumstances, the expectation would be that measure developers should have sufficient data to conduct reliability testing at the accountable entity level after three years.

The other Advisory Groups also raised concerns about rare events, which yield low data volumes. This should be less of an issue at the accountable entity level than at the patient/encounter level; nevertheless, SMP members agreed this could be a viable justification for limited testing at the accountable entity level at maintenance.

Following a thorough discussion of these potential justifications and the concerns of the other Advisory Groups, the SMP confirmed their continued recommendation for this requirement.

Validity Testing Requirements

Dr. Teigland introduced the final item of discussion. The SMP made two recommendations for validity testing at maintenance: it must be required at the accountable entity level, and it must include empirical validity testing as well. These items were also reviewed and supported by the Advisory Groups, with additional questions for the SMP.

After discussing several exceptions and best practices as outlined below, the SMP reiterated their support for the initial recommendations. Acceptable justifications for a lack of empirical accountable entity level validity testing at maintenance can include the following: (1) lack of a suitable comparator; (2) very strong Randomized Control Trial (RCT) evidence supporting a process measure and its high-fidelity process, which is unlikely to be diluted in practice or suffer from problems of generalizability in

the measured population; (3) the measure itself is the gold standard; and (4) the measure submission includes a very strong risk adjustment model to account for variation in patient characteristics, such that observed-to-expected ratios demonstrate the measure's validity. These would be situations in which the SMP's evaluation would focus on these alternative demonstrations of validity instead of on accountable entity level reliability testing. The SMP requested that measure developers continue to submit accountable entity validity testing if they are able and include additional explanation and justification if the evidence is not strong.

Regarding RCTs as evidence, not all of the SMP members agreed that this could be a strong explanatory method because evidence of the process' relationship to the outcome would not necessarily demonstrate the validity of the process as operationalized in the measure's specifications. Additionally, RCTs may not apply to the broad clinical experience, and validity testing takes place in less controlled environments.

Regarding testing when a measure is the gold standard, the SMP has previously discussed this matter. In the past, the SMP has suggested that measure developers test for a relationship between outcome variations and process variations that are conceptually related to quality of care.

In general, SMP members shared their concerns with construct validity testing, as it is usually presented. One SMP member raised concerns that an ecological fallacy could be a concern with correlation tests. Empirical validation, the way it is currently done, is weak (showing results of 0.1 or 0.2), not because a measure is poor, but because a strong correlation does not intrinsically exist between any one process and a complex outcome made up of a myriad of processes (e.g., cardiac surgery). Instead, those analyses should be done in a multilevel model with patient-level data, although other SMP members acknowledged this can be difficult to accomplish. One SMP member suggested that risk adjustment should be at the center of validity testing for adjusted outcome measures and not correlations. The SMP would like measure developers to move towards multilevel analysis in the future to better account for the possibility of an ecological fallacy. Regression-based approaches can be complex and biased.

Evaluating Testing at the Accountable Entity Level for Maintenance Measures

Dr. Teigland also introduced a new topic that the SMP did not previously discuss. During the measure evaluation meetings, SMP members have questioned how to prioritize testing presented for maintenance measures at both the patient/encounter level and accountable entity level when they present varying information of a measure's reliability or validity. NQF's measure evaluation criteria and guidance do not specify which form of testing should be prioritized.

While some members expressed interest in prioritizing the accountable entity level testing at maintenance, the SMP acknowledged that this will become more relevant once reliability thresholds become explicit and both levels of testing are required for reliability and validity. The SMP will continue discussion on this newer topic, which will grow in importance as the other recommendations from this meeting are put into effect. Future discussion should focus on whether this prioritization differs for reliability and validity testing.

Public Comment

Ms. Ingber opened the web meeting to allow for public comment. Don Casey from the American College of Medical Quality noted that as measures are used for public reporting, the nuances, disclaimers, and cautions that the SMP discussed during these meetings are not made clear to the public when viewing measures. Mr. Casey suggested that it may be helpful for the public to know this information.

Next Steps

NQF staff will conduct impact analyses on the measure portfolios for the minimum reliability testing thresholds and maintenance reliability and validity testing requirements. The next SMP advisory meeting will take place in February 2022. The SMP measure evaluation meetings for spring 2022 will take place in late March, and for fall 2022, they will take place in late October. NQF staff will send out a poll to SMP members to identify dates for 2022 meetings.