NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel May Web Meeting

The National Quality Forum (NQF) convened a public web meeting for the Scientific Methods Panel (SMP) on May 4, 2021.

Welcome, Introductions, and Review of Web Meeting Objectives

Sheri Winsper, NQF senior vice president of Quality Measurement, began by welcoming participants to the web meeting and providing opening remarks. The SMP Co-Chairs Drs. David Nerenz and Christie Teigland also shared opening remarks. Dr. Sharon Hibay, NQF senior consultant, then shared the following meeting objectives:

- Improve and clarify guidance on reliability testing for future measure development and evaluation cycles
- Discuss overarching themes that arose from previous SMP discussions
- Discuss testing considerations as related to NQF's Best Practices for Developing and Testing Risk Adjustment Models project (referred to as the "Risk Adjustment project")
- Review topics for future advisory meeting discussions

Reliability Thresholds

The SMP has discussed the issue of reliability thresholds for scientific acceptability in measure evaluations at several meetings. The foundation of this discussion topic arises from developers' frequent use of the 1977 Landis and Koch article titled *The Measurement of Observer Agreement for Categorical Data* to defend reliability testing results with a minimum threshold of 0.4. In previous meetings, SMP members agreed that the article presents arbitrary adjectives for reliability thresholds and does not provide mathematical proof for its global use in setting thresholds. They also agreed that 0.4 is too low for many situations, so the members have set out to develop clear guidance that lays out appropriate tests of reliability. Therefore, SMP members initiated the development of the *NQF Draft Acceptable Reliability Thresholds* document. The goal of the document is to provide both developers and evaluators with numerical "rule of thumb" guidance on reliability tests and adequate thresholds values for reliability results (i.e., unacceptable, acceptable, and high). *Note: This draft document was developed to guide the SMP advisory discussion and was not a final document for SMP recommendation or voting.*

The draft document organizes focal content in the reliability testing table with three categories: (1) person/encounter- (i.e., data element) level testing, (2) accountable reporting entity- (i.e., performance measure score) level testing, and (3) other reliability considerations. The table rows contain statistical approaches, tests, and potential numerical thresholds. The columns contain definitions of these tests and levels of acceptability for test purposes, ranges, and proposed threshold values. The proposed thresholds are sourced from multiple hallmark citations that are used by developers and were recommended by SMP members. A reference list will be included in a final document. The NQF Draft Acceptable Reliability Thresholds document is available for review.

SMP members reviewed the draft table and provided comments on its structure and contents, agreeing on several points. First, in general, a table such as this could be valuable for providing significant guidance to developers:

- Reliability at the person/encounter level is generally perceived as easier to obtain, and final thresholds for measures at the person/encounter level could conceivably be higher than the accountable reporting entity level.
- Additional statistical tests of reliability should be added to the table, including all reliability tests that the SMP reviews.
- The use of the terms *unacceptable*, *adequate*, and *high* is arbitrary, and they should be further defined and/or reconsidered, perhaps changed to *unacceptable* and *acceptable*.
- A finalized table should include additional information for testing appropriateness by measure type (e.g., signal-to-noise testing is generally not appropriate for measures with low-provider performance).

Other reliability testing considerations were raised by SMP members; however, agreement on these topics was not obtained. These issues and possible trade-offs should be weighed carefully in determining the most appropriate thresholds while considering strategies to fill measure gaps. Select examples include the following:

- Simple tests of correlations, as some members believe these tests are never appropriate versus other SMP members who believe these tests may be applicable based on circumstance and approach
- Defining a single “cutoff” or “threshold” for reliability testing results versus developing individual thresholds based on reliability test and approach
- Defining separate thresholds for initial endorsement and maintenance of endorsement versus maintaining consistency for both evaluation reviews
- Considering the effects of setting high thresholds that could discourage developers from submitting measures that ultimately improve care delivery and outcomes and enhance patient decision making versus the unintended consequences of unreliable measures for accountability and financial penalties. This could also have an unintended consequence of reducing available funding and healthcare access for the most vulnerable patients.

Next Steps

Although SMP members expressed general support for the table, they have many questions about its contents, and the varying member differences that still need to be considered. SMP members will continue discussing the table contents at the July web meeting. A small group of members will continue to work on the table offline and return it to the group at that time. If evaluation policy changes are recommended, they may need to undergo public commenting and CMS review, followed by review and approval from the Consensus Standards Approval Committee (CSAC). If and when the changes are enacted, they will be incorporated into the guidance and evaluation criteria. NQF often allows up to a one-year advance notice between changing criteria and implementing the changes. NQF teams will also disseminate changes, provide educational resources, and clarify changes for measure developers through various venues.

Overarching Themes

Voting on Individual Measures With Multiple Components and Performance Rates

The SMP periodically reviews individual measures that include calculations of multiple performance rates, yet no overall rate. These measures pose significant challenges to scientific reliability evaluation, including how to vote on the measures. During the March 2021 SMP measure evaluation meeting, the SMP specifically asked to discuss voting options for individual measures with multiple components and performance rates during this advisory meeting. NQF's current evaluation guidance states, "The following will not be considered composite performance measures for purposes of NQF endorsement at this time: measures with multiple measure components that are assessed for each patient, but that result in multiple scores for an accountable entity, rather than a single score. These generally should be submitted as separate measures and indicated as paired/grouped measures." (p. 51)

The SMP had questions on how to review and evaluate these measure submissions, as measures with multiple components and performance rates are often survey-based measures that include multiple survey questions or items. With the heightened desire for patient-reported outcome performance measures (PRO-PMs) that focus on patient-defined preferences, functional outcomes, and experience of care, the SMP anticipates a potential increase in these types of measures. Hence, the SMP asked to discuss evaluating these submissions and considered several options for addressing this issue in future measure evaluations:

1. Follow the above stated policy and modify the measure submission forms to align with this guidance.
2. Evaluate and vote on measures with multiple components and performance rates as an "all or nothing" measure package.
3. Assign individual NQF measures (or sub-numbers) for each measure component, and vote on individual components within the measure. NQF does not currently vote on individual components within a measure with multiple components and performance rates.

Numerous SMP members discussed the burden of developers and evaluators assembling and reviewing potentially redundant submission information in current endorsement guidance and documents when measures are paired or grouped. Some members suggested creating a new type of measure submission form specifically designed for multiple component measure sets or surveys that allow members to rate each of the measures. This would create efficiencies for developers to singularly present duplicative measure content for component measures and identify when content is dissimilar as needed. The SMP recommended that NQF maintain and adhere to the stated policy and modify the submission forms as suggested; they also recommended that measures be submitted and voted on separately. NQF staff will work through the processes of operationalizing this suggestion. SMP members and NQF staff hope this effort will reduce the burden for both developers and the reviewers by streamlining this process.

Accountable Entity/Measure Score Testing Policy

The NQF measure evaluation guidance encourages (but does not require) developers to submit accountable reporting entity- level reliability testing for maintenance of endorsement if not previously submitted in the initial endorsement. The SMP recognizes that patient/encounter level testing is often submitted for initial endorsement based on increased data element accessibility. General SMP discussion clarified that the ability to differentiate performance among and between groups is the measurement priority. Accountable reporting entity-level reliability testing requires measures to be implemented for reliability testing to be conducted, which may not have occurred by the initial endorsement. The topic was previously discussed during numerous SMP measure evaluation and

advisory meetings. During the March 2021 measure evaluation meeting, the SMP asked to discuss reaching consensus on recommending that accountable reporting entity-level reliability testing be formally required by the time of maintenance of endorsement. The SMP's concern about NQF's current policy is raised for maintenance measures undergoing maintenance evaluation with only patient/encounter-level testing, although data from implementation should be available for accountable reporting entity-level reliability testing. Further, NQF's evaluation policy allows acceptable patient/encounter-level *validity* testing to suffice for patient/encounter-level *reliability* testing. Furthermore, the current guidance and algorithm allow for measures with good patient/encounter-level testing to pass, despite having very poor accountable reporting entity-level reliability testing. Changing NQF's policy to require accountable reporting entity-level reliability testing for all maintenance measures would eliminate this loophole. All SMP members present on the call agreed with this change. No objections were voiced in response to this change.

NQF will seek and engage the measurement community for their input and perspectives on these changes. NQF will also continue to offer technical assistance to developers with attention to this potential change. Additional discussions will be needed on requiring empirical validity testing for all maintenance measures and on prioritizing accountable reporting entity-level reliability and validity testing when patient/encounter-level validity testing is also present.

Risk Adjustment Project

Dr. Matt Pickering, NQF senior director, introduced the SMP members to the NQF project on Best Practices for Developing and Testing Risk Adjustment Models. SMP members were asked to provide their input on the Technical Guidance document under development by NQF staff and the NQF-convened Risk Adjustment Technical Expert Panel (TEP). This guidance is not intended to be prescriptive about specific methods to employ but will rather present minimum standards for consideration by measure developers as they develop and test risk adjustment models that account for social and/or functional status-related risk.

Since SMP members will ultimately use the guidance to facilitate their review of risk adjustment models, NQF staff sought to gather their input on various aspects of the guidance during this web meeting. SMP members advised developers to be very thoughtful about the goals of the measure in relation to risk adjustment. Specifically, implemented models should not prevent providers from being held responsible for aspects of quality that are inherent to the measure. Conversely, when unadjusted measures are used for value-based care delivery, providers can incur financial harm when caring for populations with increased social disadvantages, which may result when safety net hospitals are compared to non-safety net hospitals.

SMP members also discussed other considerations for testing risk adjustment models. One SMP member raised a concern on the importance of requiring statistical significance of individual risk factors in deciding whether to include them in the model. This does not account for situations in which factors are correlated and therefore compete in a risk adjustment model. In these situations, individual coefficients may not appear significant because they are battling each other to explain the same variance. If developers only consider coefficients individually when testing theoretically grouped factors, they are probably not structuring their analyses properly and clearly. It is almost always impossible to meet the threshold for inclusion in the model if social risk factors must change model calibration. One SMP member pointed out that clinical risk factors do not have to meet this standard for inclusion. Instead, they are often included based on face validity. A significant change in model calibration should not be expected of social risk factors, as the ordering of risk factors may have an impact on the results of risk adjustment model testing. The clinical risk factors almost always overpower the social risk factors

when included in a model. When social risk factors are added after the inclusion of clinical risk factors, it is rare that they will have an impact on the model's calibration. However, SMP members did not feel this was a reason to eliminate them from the risk adjustment model. The C-statistic (i.e., concordance statistic or C-index), which is a measure of goodness of fit for binary outcomes in a logistic regression model, should not be the single measure of model performance presented by measure developers.

Simultaneously, one SMP member cited examples of measures evaluated by the SMP with risk adjustment for social risk factors that showed a statistically significant change in the model calibration; however, the social risk factors were not ultimately included in the final model. The SMP member stated that this is frequently performed in measure evaluations and requested that NQF provide technical guidance to address this concern. SMP members also discussed issues specific to cost and resource use measures, recommending that this be considered by the TEP; namely that risk adjustment models that include standardized pricing can hide the differences in actual resources available across providers who serve more socially disadvantaged populations. Providers that care for high-need and high-cost patients tend to have lower performance for cost and resource use quality measures. The assumption is that this is due to inefficient care delivery. However, it is not clear whether the low performance results from limited resources that are hidden by the methods or from high quality care that costs more than similar organizations with fewer high-cost and high-need patients.

Public Comment

Caitlin Flouton, NQF senior analyst, opened the web meeting to allow for public comment. One member expressed his appreciation for the SMP members' recognition of the subjectivity of the adjectives used in the Reliability Thresholds discussion.

Another member of the public asked for clarification on whether the risk adjustment framework would consider whether the measure is ultimately being used in Value-Based Purchasing (VBP) programs. NQF replied that the TEP noted that evaluation of a measure's use would be out of the purview of NQF-endorsement. This type of measure evaluation would require different criteria dependent on the intended use (i.e., evaluating validity and reliability for each use-type). However, the intent of this guidance is to provide a standard approach to social and/or functional risk adjustment within performance measurement. As such, the standards outlined are to provide developers with the necessary tools needed for NQF-endorsement, respective to social and/or functional risk adjustment.

Next Steps

Hannah Ingber, NQF senior analyst, reviewed the next steps. SMP members will be invited to participate in the upcoming Risk Adjustment web meeting on May 13, 2021. The July 2021 SMP advisory web meeting will also be rescheduled for the last week of July due to scheduling conflicts with NQF's Annual Conference. NQF will send out communications to SMP members to reschedule this meeting.