NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel – December 2020 Topical Web Meeting

The National Quality Forum (NQF) convened a public web meeting for the Scientific Methods Panel (SMP) on December 8, 2020.

Welcome, Introductions, and Review of Web Meeting Objectives

Sai Ma, NQF Managing Director and Senior Technical Expert, opened the meeting with welcoming remarks and an overview of the agenda. NQF CEO Shantanu Agrawal and Senior Vice President Sheri Winsper provided welcoming remarks as well. Additional remarks were provided by SMP Co-Chairs Drs. David Nerenz and Christie Teigland.

This meeting covered a discussion of how to continue improving the SMP review process based on the survey results of the Fall 2020 measure evaluation process, and a discussion regarding short-term improvements to NQF measure evaluation guidance and potential long-term changes to the Scientific Acceptability criteria (i.e. validity and reliability criteria).

Improving the SMP Review Process

Hannah Ingber, Senior Analyst, reviewed results of a post-meeting survey sent to the SMP members. In general, Panel members felt the workload for the cycle was as they expected and that the meeting went very well. Some specific comments suggested that for 8-9 measures, four total weeks of review time is needed.

The SMP discussed re-introducing the subgroup meetings that used to take place before the two-day measure evaluation meeting. These closed meetings could be an informal opportunity for the SMP members to confer on their interpretations of submissions and streamline the discussion items in the larger meeting. However, one SMP member argued that large group discussion is best at this point in the SMP's development for making scaled decisions. NQF will offer the opportunity for Panel members to join these informal and voluntary meetings during the next review cycle to evaluate the need for this option. Other members suggested using zipped files so that all SMP members could have materials for all measures, and another member suggested improving the flow of the evaluation form to make it follow the submission forms more closely.

Improving Guidance on Scientific Acceptability Criteria (Validity and Reliability)

Dr. Ma provided an overview of the needs for and the process on how to update the NQF evaluation guidance. She explained that NQF staff can make clarifications, minor changes, and semantic updates to the guidance during its annual upkeep, but significant and substantive changes to the evaluation criteria, processes, and policies must follow a standard multistakeholder review process. First, proposed changes must be reviewed by the SMP, then go through public commenting and CMS review, and then be reviewed and approved by the Consensus Standards Approval Committee (CSAC) for any changes to take place. Depending on the topics, proposed changes may need to be reviewed and approved by the NQF Board as well. If changes are enacted, they will be incorporated into the guidance and evaluation criteria. NQF often allows up to a 1-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.

The SMP went on to discuss the following areas regarding scientific acceptability criteria where updates are needed:

Improving Guidance on Expectations for Validity Testing and Correlation Analysis

Currently, the NQF guidance on composite measures only requests that a developer show a relationship between process, composite, and outcome, not a relationship in the "right" direction. The panel reviewed a composite measure during the fall 2020 measure evaluation cycle that showed a relationship in the "wrong" direction. Higher scores of the composite measure were associated with worse outcomes. This raised questions for the SMP about requirements for relationship directions. This measure passed largely on a technicality because the guidance does not explicitly ask for a direction, but in the future the Panel would like to see correlation analyses in a direction showing that a process measure is associated with improved outcomes or vice-versa. There was consensus among the SMP that it should be required for developers to describe the expected directional relationship between a measure and relevant outcomes so that the Panel may assess these hypothesized relationships appropriately.

As a next step, select SMP members will work with NQF staff to propose language and get the SMP's consensus on this matter in a future meeting.

Composite Measures

Current NQF guidance does not require the developer specify a "reflective" or "formative" model in their submission form for composite measures. SMP member Dr. Sherrie Kaplan led the discussion on why developers should describe which model they used to develop a composite measure, and the implications for evaluation. She explained that reflective models represent the classical concept of measurement used in psychometrics: all measures (or survey items) reflect the same underlying construct, and one therefore expects some level of correlation among the items or measures. Formative models do not assume an existing, underlying construct, but rather create the construct from the measures.[1] ("Socioeconomic Status or SES", for example, is not a single underlying construct that existed before there were measures of it; it is a creation of human analysts who use it as a way to combine multiple measures of social risk factors such as education and income.) The measures jointly determine the meaning of the construct. Therefore, for measures based on reflective model, items should be correlated and an internal consistency test (e.g., Cronbach's alpha) should be required. In contrast, for measures based on formative model, items are not correlated and therefore an internal consistency test is not appropriate. The SMP found Dr. Kaplan's overview very helpful and agreed that specific guidance should be added to the evaluation guide so that the SMP may assess whether the appropriate tests were used.

One member also pointed out that a third scenario may be considered. For example, mortality and readmissions measures are related but measure complementary outcomes. They are still evaluated as

[1] Avila, M.L., Stinson, J., Kiss, A. *et al.* A critical review of scoring options for clinical measurement tools. *BMC Res Notes* 8, 612 (2015). https://doi.org/10.1186/s13104-015-1561-6

separate measures by NQF, but they could be paired measures or balancing measures, so the question is when should they be composited vs. treated as a pair or set? SMP members also discussed the implications of asking for reflective and formative models on risk adjustment. SMP member Patrick Romano noted that for composite outcome measures, each component measure is typically risk adjusted separately, so this framework doesn't necessarily have any implications for risk adjustment.

Next step: select SMP members will work with NQF staff to propose language and get the SMP's consensus in a future meeting.

Further Clarifying Testing Requirements for Instrument-Based Measures (Including PRO-PMs)

Unlike most other measure types, instrument-based (including PRO-PM) measures are required to show tests at both data element and measure score levels for reliability and validity. However, developers often are confused by what data element and measure score refer to in the case of instrument-based measures. SMP member Zhenqiu Lin led the discussion and suggested clarifications using the terminology related to endorsement of PRO-PMs. PRO refers to patient-reported outcome, a PROM refers to an instrument used to measure the PRO, which is at the patient level, and a PRO-PM refers to a performance measure based on a PROM at the measured entity level.[2] In the context of patient-reported outcome measures, data element testing refers to the PROM (patient-level) rather than individual items of the instrument. A measure score level test (provider-level) is required for PRO-PMs. For NQF endorsement, tests of both PROM and PRO-PMs levels of tests are required. The SMP found this overview clear and pointed out that additional language needs to be more general and applicable to all instrument-based measures – not just those involving patient-reported outcomes. One SMP member stated that the higher level (e.g. PRO-PM) tests providers across patients, which by definition would show lower reliability than the lower level (e.g. PROM) that tests patients. The SMP had a discussion of what plain language is best to describe the measure score level: provider level, unit of comparison, accountable entity, measured entity, etc. Next step: select SMP members will work with NQF staff to propose language, clarifications and get the SMP's consensus at a future meeting.

Discrepancies in Evaluation Policy and Processes

Currently, under NQF evaluation guidance, for outcome, intermediate clinical outcome, cost/resource use, structure, and process measures, NQF requires EITHER data element testing OR score-level testing for both reliability and validity. For new measures, face validity testing of the measure score is accepted. At the time of maintenance, empirical testing of validity is expected. The level of testing does impact potential evaluation ratings: a measure with only data-element-level testing can only get a moderate rating, at best. A submission that presents measure-score-level testing can receive a high rating.

During the last review cycle, the SMP members pointed out that it is unclear whether and why these policies should apply to maintenance measures, since maintenance measures are supposedly already in use for measuring performance of accountable entities and should have access to real-world test data. The SMP discussed whether score-level testing should always be required of maintenance measures and whether empirical testing of validity should always be required of maintenance measures.

The SMP in general agreed it is difficult to make a positive evaluation of a measure without measure score level reliability and validity testing. A few SMP members stated that the real question is whether NQF wants to require data element validity and reliability testing if measure score level testing is

[2] NQF Report (2013): Patient Reported Outcomes (PROs) in Performance Measurement https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=72537

conducted and results are adequate. CMS measures, for example, are usually based on administrative data and claims. Although these codes can generally be regarded as reliable, validating and auditing these data requires a large investment in resources. The Panel must consider the burden for smaller developers who use other non-CMS data sources in order to balance expectations. In other words, the resources needed for them to validate the reliability and validity of their data sources needs to be considered. Another SMP member cautioned that even CMS data elements can be affected by billing policy. For example, safety net providers with low Medicare caseloads do not have financial incentives to code comorbidities as comprehensively as other providers but it would affect them if those comorbidities are used in developing measures or in risk adjustment. In relation to this issue, an SMP member used BMI as an example to illustrate data elements (weight and height) that need to be accurate for the measure (BMI) to be reliable and valid.

Although the SMP members agreed that for maintenance measures, measure score-level tests should be expected, they wondered whether there might be legitimate reasons and circumstances why developers of maintenance measures could not provide such tests (e.g. not enough test sites).

Next step: NQF staff will review past submissions and identify reasons as to why developers could not provide empirical tests for validity, or tests at the measure score level. NQF staff will present the findings in the future meetings and propose changes.

Varying Scientific Acceptability Criteria by the Purpose of Measure Use

Currently measures are reviewed for scientific acceptability (validity and reliability) regardless of the future potential measure use. However, in the past few review cycles, the SMP members stated that they could not accurately assess validity and reliability of a measure without knowing the intended use. For example, a measure used for identifying outliers does not need to meet the same reliability standards as a measure used for payment purpose such as a 5-star program. The questions for the SMP to consider are: Should evaluation criteria vary according to the measure purpose categories? If so, how? Should measures be endorsed for only specific uses for which the testing was completed?

Co-Chair Dr. Nerenz led the discussion by offering the following categories of uses of measures as a way to frame the discussion:

1. Identify outliers
2. Put entities into one of two groups to either get or not get financial reward or punishment (e.g., lowest 25th percentile)
3. Group entities into quintiles or similar groupings for "star ratings"
4. Group entities into deciles for purposes of financial rewards or punishments
5. Support consumer choice among similar entities in a geographic or market area (this would often involve making distinctions among "three-star" entities
6. Use continuous scores for either consumer choice or financial incentive payments

There was a general agreement that reliability and validity should be evaluated through the lens of use or purpose but there was no agreement on how to categorize the uses. Several SMP members pointed out that some star programs do not show a wide distribution of scores, indicating that there is not a big difference between each star level. Other members also mentioned that it should be clear that the SMP is not attempting to adjudicate or set thresholds for those programs. A few SMP members expressed the need for confidence intervals and/or standard deviations for statistics in the measure submission.

Next step: select SMP members will work with NQF staff to propose language and get the SMP's consensus in the future meetings.

Adding Settings

Currently the NQF measure submission form lists settings as individual clinician, group/practice (hospital/facility/agency), health plan, and other. However, NQF has observed an increased number of measures submitted for levels such as ACOs, National Provider Identifiers (NPIs), hospital unit/department, etc. The questions for the SMP to consider include: Should other levels be added and what are the concerns? (e.g. NPI can range from an individual doctor to an organization (group practice, lab, hospital, etc.). If we allow for a level with mixed units, what guidance can we offer?

The SMP in general agreed that ACO could be added explicitly as a level to the submission form. However, several SMP members suggest not to mix identifiers with a level of measurement. More discussion of these issues is needed.

Next step: select SMP members will work with NQF staff to propose language and get the SMP's consensus in the future meetings.

Acceptable Thresholds for Reliability

During the last measure evaluation meeting, the SMP agreed that the Landis & Koch "scale"[3] was arbitrary to begin with and not applicable anymore and NQF would like to provide stronger and more specific guidance for developers. The questions for the SMP to consider include: What are acceptable alternatives to Landis & Koch? What conceptual approaches are acceptable without an appropriate threshold? In lieu of the Landis & Koch paper or Adams tutorial[4], what guidance and references can we provide to developers?

The SMP generally agreed that they do not want to suggest a different set of arbitrary thresholds (e.g. Koo and Li (2016)[5] thresholds are also arbitrary). They also pointed out this discussion is linked to the earlier conversation about applying the lens of intended use.

A few ideas were discussed: a couple of members suggested that at least have 0.25 variance (squaring the reliability coefficient of 0.50) in the measure should be reliable. Several members stated that stability and risk of misclassification are the main concerns here. If a developer can complement their reliability tests with an analysis of risks for misclassification (i.e. how entities will change their rankings through a simulation), those concerns could be alleviated. However, the SMP members recognize this requires a great deal of resources for the developers.

NQF staff asked whether the SMP can suggest a minimum acceptable threshold for reliability tests, or maybe use pass/fail instead of rating, while working on a larger policy change. Nearly all SMP members agreed that 0.4 for reliability tests is too low. It was suggested that 0.5 or 0.6 may be acceptable as a minimum acceptable threshold for the time being.

Risk Adjustment

A question about risk adjustment was included in the slide deck: whether a measure can be rated "insufficient" solely due to a lack of or inappropriate use of risk adjustment? Given the limited time, this topic has been moved to a future meeting for discussion.

---

[3] *Landis J, Koch G. The measurement of observer agreement for categorical data, Biometrics 1977;33:159-174.*
[4] *Estimating Reliability and Misclassification in Physician Profiling. John L. Adams, Ateev Mehrotra, Elizabeth A. McGlynn, RAND 2010*
[5] *Koo & Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research J Chiropr Med. 2016*

Public Comment

Dr. Ma opened the web meeting to allow for public comment. No public comments were offered.

Next Steps

Caitlin Flouton, NQF Analyst, summarized the next steps for the SMP: this meeting summary will be circulated for review by the SMP members. NQF Staff will send out a poll to collect availability of SMP members for the upcoming year, which is due back by December 14. The SMP evaluation meeting for the Spring 2021 cycle is likely to be at the end of March 2021.